Prédictions bioinformatiques des propriétés des domaines de reconnaissance peptidique.

Emmanuelle Becker

Ecole Doctorale Interface de la Physique, de la Chimie et de l'Informatique avec la Biologie
Université Paris 6 – Pierre et Marie Curie

Thèse présentée pour obtenir le grade de Docteur en Sciences de l'Université Paris 6 par

Emmanuelle BECKER

Prédictions bioinformatiques des propriétés des domaines de reconnaissance peptidique.

Thèse dirigée par Raphaël GUEROIS

à soutenir le 26 Septembre 2007 devant le jury composé de :

Madame Anne POUPON (Chargée de recherche CNRS à l'IBBMC, Orsay - Rapporteur)
Monsieur Gilles LABESSE (Chargé de recherche CNRS au CBS, Montpellier - Rapporteur)
Monsieur Olivier LEQUIN (Professeur à l'Université Paris 6, Paris - Président du jury)
Monsieur Raphaël GUEROIS (Chercheur CEA à l'iBiTecS, Saclay - Examinateur)
Monsieur Jean-Michel NEUMANN (Chercheur CEA à l'iBiTecS, Saclay - Directeur de thèse)

Liste des Abréviations

ADN	=	Acide Désoxyribo-Nucléique
AP-MS	=	Affinity Purification – Mass Spectrometry
Asf1	=	Anti-Silencing Function 1
BBDEP	=	Backbone Dependant (fr. dépendant du squelette peptidique)
BBIND	=	Backbone Independant (fr. indépendant du squelette peptidique)
BIND	=	Biomolecular Interaction Network Database
BRCT	=	Breast Cancer carboxy-Terminal
CASP	=	Critcal Assessment of protein Structure Prediction
Da	=	Dalton
DIP	=	Database of Interacting Proteins
DSB	=	Double Strand Break (fr. cassure double brin)
EGTA	=	acide Ethylène Glycol TétraAcétique
EMBL	=	European Molecular Biology Laboratory
FHA	=	ForkHead Associated
GST	=	Glutathion S Transferase
HMM	=	Hidden Markov Model (fr. modèle de Markov caché)
IgG	=	Immuno-globuline G
IPP	=	Interaction Proteine-Proteine
Kd	=	Constante de Dissociation
M	=	Molaire
MS	=	Mass Spectroscopy (fr. spectrométrie de masse)
NBS	=	Nijmengen Breackage Syndrome (fr. syndrome de Nimègue)
ORF	=	Open Reading Frame (fr. Cadre ouvert de lecture)
PDB	=	Protein Data Bank
PPI	=	Protein Protein Interaction (fr. interaction protéine-protéine)
PRM	=	Peptide Recognition Module
PSSM	=	Position Specific Scoring Matrix
pS	=	phospho-sérine
pT	=	phospho-thréonine
RMN	=	Résonance Magnétique Nucléaire

Abréviations.

RMSD	=	Root Mean Square Deviation (fr. écart quadratique moyen)
RRM	=	RNA Recognition Module
SCMFT	=	Self Consistent Mean Field Theory
SCOP	=	Structural Classification Of Proteins
SCP	=	Side Chain Positionning
SCWRL	=	Side Chain Weighted Rotamer Library
SGD	=	Saccharomyces Genome Database
SSDEP	=	Secondary Structure Dependant (fr. dépendant des structures secondaires)
TAP	=	Tandem Affinity Purification
TEV	=	Tobacco Etch Virus
Y2H	=	Yeast double hybrid.

Sommaire

Chapitre 1 : Introduction Générale

La vision du fonctionnement cellulaire a été profondément bouleversée ces dernières années par l'essor des analyses protéomiques à grande échelle. Les premières cartes globales d'interactions protéine-protéine révèlent l'existence de réseaux d'interactions fortement interconnectés dans lesquels la grande majorité des protéines se trouvent à proximité les unes des autres. Cette complexité soulève un grand nombre de questions quant aux stratégies cellulaires permettant d'orchestrer l'activation de ces différents réseaux. Les mécanismes de réparation de l'ADN sont à ce titre remarquables car ils requièrent l'activation concertée de plusieurs voies de signalisation très différentes (Rouse and Jackson, 2002). Ils impliquent en effet des fonctions aussi variées que la détection des lésions, l'activation de la transcription de gènes spécifiques ou le blocage transitoire du cycle cellulaire à différentes étapes (Bartek et al., 2001).

Les phénotypes induits par la délétion de grandes régions ou de la totalité d'un gène permettent de mesurer son importance dans le processus de réparation. Néanmoins, de telles modifications provoquent généralement des perturbations trop drastiques pour interpréter le rôle fonctionnel des différentes interactions dans lesquelles une protéine est impliquée. La dissection des réseaux d'interaction et la compréhension de leur coordination interne impose de les perturber plus localement en abrogeant ou atténuant une interaction spécifique au sein du réseau. Cette phase de l'analyse fonctionnelle suppose la connaissance préalable des sites d'interaction, qui constitue au plan général un enjeu majeur pour les méthodes bioinformatiques.

Un élément clé pour appréhender la complexité des réseaux d'interactions est l'existence de domaines médiateurs d'interactions spécialisés dans le recrutement spécifique de partenaires cellulaires. Ces domaines, également appelés modules de reconnaissance peptidique (PRM), jouent un rôle fondamental dans la logique moléculaire associée aux réseaux d'interactions. Prédire par une analyse bioinformatique les sites reconnus par ces PRMs constitue un enjeu majeur au plan biologique et justifie le développement de nouvelles méthodologies bioinformatiques au cœur de ce travail de thèse.

Dans l'introduction qui suit, nous présentons tout d'abord le contexte biologique qui nous a mené à l'étude de ces domaines protéiques et exposons les propriétés qui les caractérisent. Par la suite, nous introduisons les méthodes bioinformatiques existantes que nous avons utilisées et/ou cherché à améliorer dans le cadre de cette thèse.

Les interactions protéine-protéine (IPP)

1.1.1 Introduction.

Le séquençage de génomes complets chez les eucaryotes a permis d'accéder à une quantité importante d'informations jusqu'alors très incomplètes. La bioinformatique a joué un rôle prépondérant dans cette avancée puisque ses développements ont largement contribué à l'analyse à grande échelle des données du séquençage pour identifier les gènes codés et leurs variants d'épissage. Ainsi, il a pu être mis en évidence que le génome de *Saccharomyces cerevisiae* code pour environ 6300 protéines alors que le génome humain en compterait 30000 (Goffeau et al., 1996; Lander et al., 2001). En plus d'un nombre plus important de gènes codés, le génome humain se caractérise également par une utilisation plus fréquente de variants d'épissage.

La question de la fonction cellulaire de ces protéines est étroitement liée à leur composition en domaines structuraux (unités autonomes de repliement), à leurs structures et à l'organisation des interactions intra- et inter-moléculaires entre leurs différents domaines.

De nombreux outils bioinformatiques ont été introduits pour faciliter l'étude du rôle de ces protéines. Tous sont basés sur le principe suivant : les séquences présentant des similitudes importantes correspondent à des protéines qui adoptent un repliement tridimensionnel proche avec des fonctions voisines. Néanmoins, des séquences très différentes peuvent aussi conduire à des repliements proches : la détection de ces similitudes structurales au travers des séquences s'avère alors bien plus délicate. Différentes techniques ont été développées pour repérer les séquences susceptibles d'adopter un même repliement ainsi que pour modéliser le repliement d'une protéine de structure inconnue. Nous étudierons plus en détails ces méthodes bioinformatiques dans la suite de l'introduction.

Pour aboutir à une vision plus intégrée du fonctionnement des cellules, les liens entre ces protéines restent à mettre en évidence. Les nouveaux défis de l'ère post-génomique se concentrent donc sur (i) la mise en évidence des interactions entre protéines et l'identification du rôle précis de chacune d'entre elles et (ii) la compréhension des mécanismes contrôlant la transcription des différents gènes.

1.1.2 Mise en évidence des interactions protéine-protéine (IPP).

Un certain nombre d'interactions protéine-protéine sont connues et référencées dans la littérature. A ces interactions connues peuvent être ajoutées des interactions observées lors de cribles à grande échelle basés sur :

- la technique du double hybride (Y2H) ;
- ou celle de la purification par affinité (AP) couplée à la spectrométrie de masse (MS).

En raison de méthodologies différentes, ces deux techniques de criblage à grande échelle identifient souvent des interactions non redondantes et s'avèrent donc complémentaires.

La technique du double hybride (Fields and Song, 1989) a été introduite afin de mettre en évidence une interactions entre deux partenaires X et Y chez la levure *Saccharomyces cerevisiae* (**figure 1**). La méthode consiste à utiliser le facteur de transcription Gal4, dont les domaines d'association à l'ADN et d'activation de la transcription peuvent être dissociés afin de construire deux protéines hybrides qui sont introduites dans des souches de levure : la protéine X est fusionnée au domaine Gal4 1-147 de liaison à l'ADN, tandis que la protéine Y est fusionnée au domaine Gal4 768-881 activateur de la transcription. Dans les souches où l'interaction entre X et Y est présente, la proximité des domaines de liaison à l'ADN et d'activation de la transcription de Gal4 permet de déclencher l'expression des gènes sous contrôle du promoteur Gal4. L'expression de ces « gènes rapporteurs » sert de marqueur pour identifier une interaction entre X et Y.

La technique du double hydride a l'avantage de mettre en évidence les interactions protéine-protéine dans un contexte cellulaire. Ainsi, si l'interaction étudiée met en jeu deux protéines de levure et nécessite une modification post-traductionnelle, celle-ci pourra être détectée par double-hybride. Néanmoins, parmi les inconvénients de la technique figurent son manque de sensibilité et de spécificité, puisqu'elle génère de nombreux faux positifs et faux négatifs. Des cribles à grande échelle utilisant cette technique ont été effectués chez la levure *Saccharomyces cerevisiae* (Ito et al., 2001; Uetz et al., 2000), le ver *Caenorhabditis elegans* (Li et al., 2004), chez *Helicobacter pylori* (Rain et al., 2001), la mouche *Drosophila melanogaster* (Formstecher et al., 2005; Giot et al., 2003) et plus récemment chez l'homme (Lim et al., 2006; Rual et al., 2005).

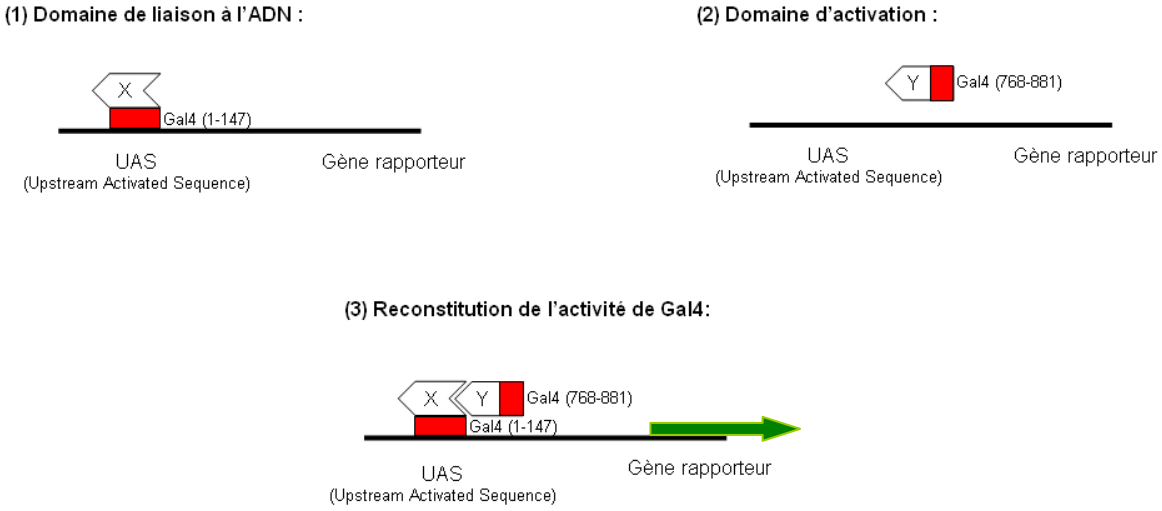


figure 1 : Principe du double hydride chez la levure *Saccharomyces cerevisiae*. (1) La protéine X est fusionnée au domaine de liaison à l'ADN de la protéine Gal4. (2) La protéine Y est fusionnée au domaine activateur de la transcription de Gal4. (3) L'interaction de X et Y permet la reconstitution de l'activité de la protéine Gal4 et ainsi l'expression des gènes sous contrôle du promoteur Gal4, considérés comme les gènes rapporteurs.

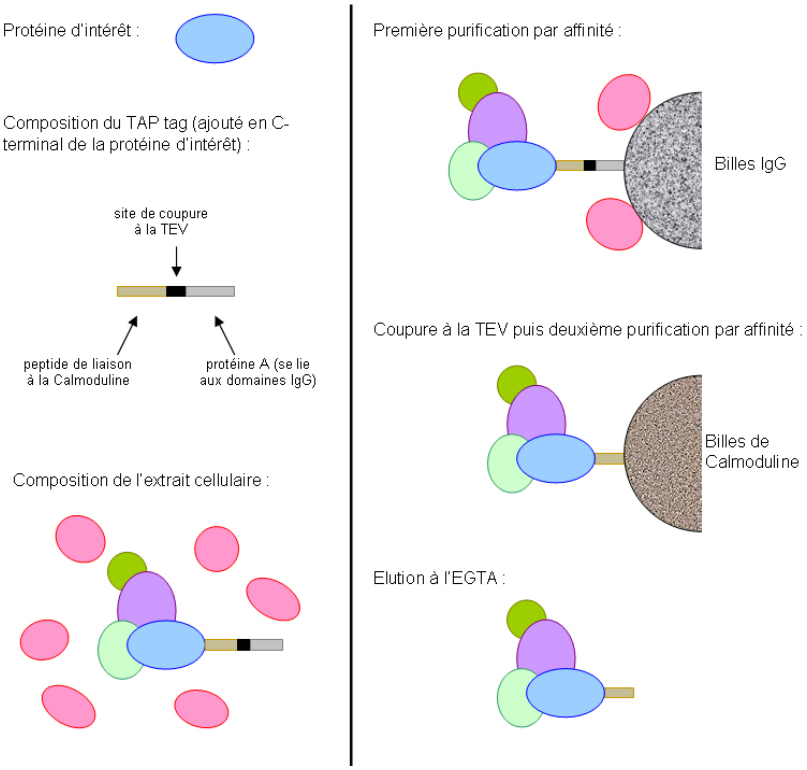


figure 2 : Processus de purification par affinité développé dans l'optique de cribles à grande échelle identifiant les composants de complexes protéiques chez *S. cerevisiae*, d'après (Puig et al., 2001). Une extension C-terminale est ajoutée à la protéine d'intérêt, comprenant un domaine de liaison aux billes IgG, un site de coupure à la TEV et un domaine de liaison aux billes de Calmoduline. Après avoir introduit le gène modifié dans les cellules de levure et cultivé celles-ci, le jus cellulaire est extrait. Il est purifié une première fois sur une colonne de billes IgG, puis après coupure à la TEV, il est purifié une seconde fois sur des billes de Calmoduline. Après élution à l'EGTA, on peut envisager d'identifier les différents composants des complexes par spectrométrie de masse.

Une autre méthode basée sur l'identification des composants de complexes protéiques par spectrométrie de masse a également été mise au point. Pour chaque protéine d'intérêt X, une extension N- ou C-terminale contenant les domaines de liaisons nécessaires à deux purifications par affinité successives est ajoutée et la construction est introduite dans la cellule ou l'organisme hôte. Des extraits cellulaires sont ensuite préparés, au sein desquels la protéine d'intérêt X est complexée à un certain nombre de partenaires non identifiés. Le complexe est purifié (**figure 2**) pour que les protéines composant ce complexe puissent être caractérisées par spectrométrie de masse (Puig et al., 2001) ou par la détection d'anticorps. Trois cribles utilisant cette approche à grande échelle ont été réalisés chez *Saccharomyces cerevisiae* (Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006).

Les informations mises en évidence par les approches Y2H et AP-MS ne sont pas équivalentes. Les interactions détectées par Y2H sont binaires (**figure 3**-B), alors que les complexes identifiés par AP-MS sont caractérisés de façon globale sans connaître précisément l'identité des partenaires en contact (**figure 3**-B). Ces deux visions sont complémentaires et leur association est utile pour inférer la structure locale exacte de la carte des interactions protéine-protéine (**figure 3**-C) (Scholtens and Gentleman, 2004; Scholtens et al., 2005).

Le cas de *Saccharomyces cerevisiae* est intéressant car plusieurs cribles à grande échelle ont été effectués, à la fois par Y2H (Ito et al., 2001; Uetz et al., 2000) et par AP-MS (Gavin et al., 2002; Ho et al., 2002; Krogan et al., 2006). La comparaison des résultats obtenus est surprenante puisqu'on constate un taux recouvrement extrêmement faible entre les interactions identifiées. Par exemple, Nervan Krogan et ses collègues (*University of Toronto*, Canada) ont comparé les résultats de leur crible AP-MS à ceux des deux cribles Y2H : sur les 547 complexes détectés par leur crible, seuls 47 sont complètement identifiés par le crible AP-MS de Yuen Ho et coll. et 52 par le crible AP-MS d'Anne-Claude Gavin et coll. (Krogan et al., 2006). Ce manque de recouvrement peut s'expliquer en partie par des différences méthodologiques expérimentales car les protéines utilisées comme proies sont différentes dans les trois cribles.

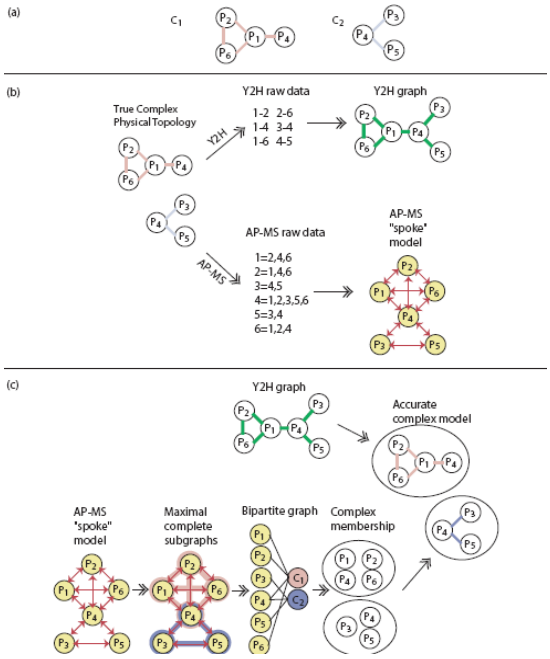


figure 3 : Exemple des interactions identifiées par Y2H et AP-MS et de leur combinaison pour retrouver la topologie locale exacte. **(a)** Topologie de deux complexes protéiques *in vivo* dans lesquels la protéine P4 intervient. **(b)** Interactions identifiées par les deux approches, Y2H et AP-MS, et les réseaux que l'on peut inférer à partir de ces résultats (réseau d'après les données de Y2H en vert, réseau « spoke » d'après AP-MS en rouge). **(c)** Combinaison des deux approches pour identifier la topologie des complexes *in vivo*.

Pour représenter toutes les IPP identifiées à l'échelle d'un organisme, la structure de données la plus communément utilisée est celle des cartes d'IPP, ou interactomes : il s'agit d'un graphe au sein duquel chaque protéine correspond à un sommet et chaque interaction identifiée à une arête. Les interactomes forment des réseaux de très grandes taille et complexité qui fournissent une nouvelle vision intégrée du fonctionnement de la cellule et de l'organisme. Leur utilisation a par exemple conduit à des progrès importants concernant la classification et l'annotation fonctionnelles des protéines (Brun et al., 2003). L'étude des propriétés statistiques de ces réseaux a suscité beaucoup d'intérêt ces dernières années. De nombreuses questions restent ouvertes concernant par exemple les modèles représentant le mieux la topologie observée (Barabasi and Albert, 1999; Przulj et al., 2004; Ravasz et al., 2002; Watts and Strogatz, 1998). En effet, au sein des interactomes, il a été constaté que si la majorité des protéines ont peu de partenaires d'interactions, d'autres au contraire sont impliquées dans un grand nombre d'interactions (ces dernières étant surnommées « *hubs* »[*]).

[*] Le terme de « *hub* » est également utilisé pour décrire les plates-formes de correspondance du réseau aérien (les grands aéroports), ou pour les concentrateurs à partir desquels on peut construire un réseau informatique local.

1.1.3 Les bases de données d'interactions protéine-protéine.

Il existe différentes bases de données d'interactions protéiques, construites à partir de données expérimentales. En particulier, les bases de données BIND (*Biomolecular Interaction Database*) et DIP (*Database of Interacting Proteins*) collectent les interactions protéine-protéine provenant de différentes sources : soumissions directes, données collectées provenant des différentes expériences menées à grande échelle, ou encore analyses manuelles ou automatiques de données issues de la littérature. A l'heure actuelle, la base de données BIND contient 67739 interactions et la base de données DIP 56080 interactions mettant en jeu 19378 protéines.

Les différentes expériences menées à grande échelle, en particulier chez *Saccharomyces cerevisiae*, *Caenorhabditis elegans, Helicobacter pylori*, *Drosophila* et chez l'homme ont largement contribué au développement de ces bases de données d'interactions protéiques. En 2003, la proportion de ce type de données dans BIND a été estimée à environ 80% (Salwinski and Eisenberg, 2003). Ces données d'interactions proviennent en partie d'expériences de double hybride qui génèrent une part non négligeable de faux positifs. Compte tenu du nombre de données à analyser, certaines méthodes automatiques ont été développées afin d'évaluer la validité biologique de ces interactions, comme les méthodes *Expression Profile Reliability* et *Paralogous Verification Method* (Deane et al., 2002). À partir de ces méthodes d'évaluation, il a été estimé qu'environ 50% des 8000 interactions mises en évidence par un crible double hybride chez *Saccharomyces cerevisiae* étaient valides (Deane et al., 2002). Les interactions prédites comme valides ont été regroupées en un sous-ensemble représentant environ 30% des interactions de la base de données DIP, appelé CORE (Salwinski and Eisenberg, 2003).

D'autres approches ont été développées afin d'enrichir le contenu des bases de données d'interactions protéiques. En particulier, certaines approches automatiques basées sur des techniques de fouilles de données visent à extraire les articles scientifiques traitants des interactions protéiques caractérisées expérimentalement (Donaldson et al., 2003; Marcotte et al., 2001). Cependant, une analyse manuelle est nécessaire dans ce cas avant d'intégrer ces données au sein des bases d'interactions protéiques.

Les bases de données BIND et DIP ont récemment intégré les données structurales concernant certains complexes protéiques connus, à partir de la *Protein Data Bank* (PDB). La *Protein Data Bank* est une base de données qui référence les coordonnées atomiques de l'ensemble des structures tridimensionnelles expérimentales et des modèles théoriques des macromolécules biologiques (Berman et al., 2000).

À partir des données structurales issues de la PDB, la base de données PSIMAP (*Protein Structural Interactome MAP*) a été développée dans le but de reconstruire une carte globale d'interactions qui décrit précisément les interactions entre domaines protéiques, aussi bien intra- qu'inter-moléculaires. L'originalité de cette base consiste à considérer la définition SCOP (*Structural Classification Of Proteins*) des domaines protéiques, qui exploite des homologies structurales et fonctionnelles pour définir des familles ou superfamilles de domaines d'origine évolutive distincte. Ainsi, PSIMAP est la première base de données à proposer une vision globale des interactions entre domaines protéiques à l'échelle de la superfamille. À l'heure actuelle, elle répertorie 1930 superfamilles SCOP impliquées dans des interactions protéiques.

D'autres bases de données sont également issues des données structurales d'interactions protéiques, et répertorient certaines propriétés physico-chimiques des interfaces protéiques comme la base de données SPIN-PP ou encore les propriétés énergétiques des interfaces avec les bases de données ASEdb ou BID.

Enfin, une initiative récente propose de référencer au sein de la base de données INTACT, librement accessible, les interactions protéine-protéine décrites dans la littérature, ou soumises directement par les utilisateurs d'INTACT (Hermjakob et al., 2004). Dans sa version de mai 2007, plus de 55000 protéines et 85000 interactions y sont disponibles.

Base de données	Description	URL
DIP (Database of Interacting Proteins)	Interactions entre protéines	http://dip.doe-mbi.ucla.edu
BIND (Biomolecular Interaction Network Database)	Interactions entre biomolécules	http://www.bind.ca/
PDB (Protein Data Bank)	Structure atomique des protéines (dont certains complexes protéiques)	http://www.rcsb.org/pdb/
PSIMAP (Protein Structural Interactome MAP)	Structure des interactions entre domaines protéiques	http://psibase.kobic.re.kr/
PQS (Protein Quaternary Quaternary Structures)	Structure quaternaire des protéines	http://pqs.ebi.ac.uk/
SPIN-PP (Surface Properties of Interfaces— Protein Protein Interfaces)	Interfaces protéiques (propriétés physico-chimie)	http://honiglab.cpmc.columbia.edu/SPIN/main.html
ASEdb (Alanine Scanning Energetics database)	Interfaces protéiques (propriétés énergétiques)	http://www.asedb.org
BID (Binding Interface Database)	Interfaces protéiques (propriétés énergétique)	http://tsailab.tamu.edu/BID/
IntAct	Interactions entre biomolécules	http:// www.ebi.ac.uk/intact/

table 1 : Bases de données d'interactions protéine-protéine. La première colonne indique le nom des bases de données, la seconde leur contenu exact et la dernière colonne l'adresse du portail de leur site Internet.

1.1.4 Les différents types d'interactions protéine-protéine.

Au sein de la cellule, les interactions protéine-protéine assurent des fonctions diverses. Trois caractéristiques principales permettent de décrire les IPP (Nooren and Thornton, 2003) :

(i) les complexes formés de chaînes identiques (homo-oligomères), et les complexes formés de chaînes différentes (hétéro-oligomères).

(ii) les complexes formés de protomères dont la structure non-complexée n'est pas stable. Ils représentent une part minime des IPP qui pourrait néanmoins être sous-estimée par les difficultés expérimentales liées à leur étude.

(iii) la nature permanente ou transitoire d'une interaction protéine-protéine. On différencie de cette manière les protéines s'associant et se dissociant de façon labile (interaction transitoire) et les protéines dont seule la forme complexée est présente (interaction permanente).

Le contexte cellulaire est primordial dans le cadre de l'étude des interactions protéine-protéine transitoires, puisqu'il peut permettre de moduler l'équilibre oligomérique entre forme complexée et non complexée. Plus précisément, dans le cas de deux protéines interagissant de façon transitoire, l'état d'oligomérisation des protéines en solution dépend avant tout de la concentration des deux partenaires et de la constante de dissociation (Kd) de l'interaction. L'environnement physiologique peut déplacer cet équilibre oligomérique par une variation de pH ou de température, par le changement de concentration d'un partenaire, d'ions ou de substrats, ou enfin par une modification post-traductionnelle comme une phosphorylation.

La présence de nombreux complexes structuraux au sein de la PROTEIN DATA BANK a permis d'étudier les caractéristiques structurales des interfaces d'interactions protéine-protéine (Argos, 1988; Janin et al., 1988; Jones and Thornton, 1995; Miller et al., 1987). Une modification conformationnelle de l'un ou des deux partenaire(s) lors de leur association est fréquente lorsque la taille de l'interface du complexe est supérieure à 1000A^2 (Lo Conte et al., 1999; Nooren and Thornton, 2003).

Dans le cas des homo-dimères, une faible corrélation a été constatée entre l'énergie de liaison ΔG et l'hydrophobicité des interfaces. De façon surprenante, cette propriété n'est pas

retrouvée pour les hétéro-dimères, pour lesquels l'énergie de liaison ΔG n'est corrélée ni à la polarité, ni à la taille de l'interface (Brooijmans et al., 2002; Nooren and Thornton, 2003).

1.2 Les domaines médiateurs d'interactions protéine-protéine.

1.2.1 Définition des Peptide Recognition Modules (PRMs).

Au sein de la cellule, la transduction des signaux est souvent médiée par des interactions protéine-protéine dans le contexte de grands complexes multi-moléculaires. La formation de ces complexes doit être régulée au niveau temporel, afin que la cellule puisse répondre à un *stimulus* selon un processus dynamique, mais également au niveau spatial, de façon à ce que la réponse puisse être dirigée vers le compartiment sub-cellulaire adéquat.

Certains domaines protéiques sont spécialisés dans la régulation des complexes multi-moléculaires. C'est le cas des *Peptide Recognition Modules* (PRMs), qui partagent comme caractéristique commune de reconnaître spécifiquement de courts fragments protéiques, souvent associés à une modification post-traductionnelle. La **table 2** présente les familles de PRMs les plus abondantes à ce jour. Beaucoup de familles de PRMs reconnaissent des résidus phosphorylés, acétylés, méthylés etc… En effet, la nature transitoire de ces modifications post-traductionnelles permet de mettre en place des voies de signalisation réversibles.

Depuis le séquençage de différents génomes, il est possible d'estimer le nombre de PRMs et leur distribution dans différents organismes. La **table 2** reporte l'abondance estimée de ces PRMs chez *Saccharomyces cerevisiae* et chez *Homo sapiens*. On constate que les PRMs sont beaucoup plus abondants chez l'homme que chez la levure, malgré la surestimation probable du nombre de PRMs de la levure (voir légende). Parmi les exemples les plus marquants, le nombre estimé de domaines PDZ passe de 3 dans la levure à 356 chez l'homme et celui des domaines SH3 de 27 à 409.

PRM	Motifs protéiques reconnus	Abondance chez *S. cerevisiae* et *H. Sapiens*	Indentité de séquence.	Image dans figure 5
WW	motifs riches en prolines [$]	9 chez *SC.* et 125 chez *HS.*	36%	
GYF	motifs riches en prolines	3 chez *SC.* et 4 chez *HS.*	29%	
SH3	motifs riches en prolines	27 chez *SC.* et 409 chez *HS.*	27%	
WH1 ou EVH1	motifs riches en prolines	1 chez *SC.* et 13 chez *HS.*	28%	
FHA	thréonines phosphorylées	14 chez *SC.* et 29 chez *HS.*	20%	
BRCT	sérines phosphorylées	18 chez *SC.* et 56 chez *HS.*	13%	F
Polo-Box	sérines/thréonines phosphorylées	2 chez *SC.* et 8 chez *HS.*	26%	
FF	sérines phosphorylées	5 chez *SC.* et 29 chez *HS.*	21%	
14-3-3	sérines/thréonines phosphorylées	3 chez *SC.* et 8 chez *HS.*	52%	E
Répétitions WD	sérines/thréonines phosphorylées [£]	104 chez SC. et 462 chez HS.	22%	C
Tudor	lysines méthylées	2 chez *SC.* et 80 chez *HS.*	17%	B
Chromo	lysines méthylées	5 chez *SC.* et 47 chez *HS.*	23%	A
PDZ	extrémités C-terminales	3 chez *SC.* et 356 chez *HS.*	19%	
PTB	tyrosines phosphorylées*	0 chez *SC.* et 61 chez *HS.*	35%	
SH2	tyrosines phosphorylées	1 chez *SC.* et 139 chez *HS.*	26%	
Bromo	lysines acétylées	14 chez *SC.* et 99 chez *HS.*	29%	D

[$] peut aussi reconnaître des sérines/thréonines phosphorylées associées à des prolines.
[£] peut aussi reconnaître des lysines méthylées

table 2 : Table récapitulative des PRMs les plus abondants. Les PRMs sont regroupés en fonction du type de fragment protéique qu'ils reconnaissent : les motifs riches en prolines (en vert), les phospho-sérines et phospho-thréonines (en bleu), les phospho-tyrosines (en rose), les lysines méthylées (en gris), les lysines acétylées (en orange) et les extrémités C-terminales (en jaune). Lorsqu'il existe des exceptions à la règle générale, celles-ci sont indiquées par des caractères spéciaux (*[$]). La troisième colonne indique l'abondance relative des différents PRMs chez *Saccharomyces cerevisiae* et chez *Homo sapiens*. L'abondance des PRMs chez *Cerevisiae* a été estimée via une interrogation directe de la banque SwissProt/Trembl. Etant donné que cette banque est très redondante, il est probable que ces chiffres soient sur-estimés (Castagnoli et al., 2004). Pour l'estimation de l'abondance des différents PRMs chez l'homme, la banque de données International Protein Index, très peu redondante, a été utilisée. La quatrième colonne indique l'identité de séquence moyenne observée au sein de la famille, calculée à partir d'un alignement local des domaines. Enfin, la dernière colonne fait référence aux structures illustrées dans les figures de ce manuscrit.

Le taux de conservation des PRMs est variable, bien qu'on puisse considérer que les PRMs ont en règle générale des séquences plutôt divergentes. En effet, parmi les PRMs les plus abondants présentés au sein de la **table 2**, le taux de conservation maximal au sein d'une famille n'est que de 52% ! Certains domaines, comme les domaines WW, ont un taux d'identité suffisant pour être efficacement détectés et alignés par des méthodes bioinformatiques automatiques (>30%). C'est également le cas des domaines PTB et 14-3-3. D'autres PRMs, avec une identité de séquence au sein de la famille comprise entre 20% et 30%, peuvent être détectés par les méthodes bioinformatiques automatiques mais il peut être délicat de déterminer avec précision les délimitations N- et C-terminales des domaines

comme d'aligner leurs séquences avec la séquence d'autres membres de leurs familles. Les domaines GYF, SH2, SH3, Chromo, Bromo, FF, WH1, Polo-Box et WD peuvent être classés dans cette catégorie. Enfin, pour un certain nombre de PRMs, la divergence des séquences est telle qu'il est parfois délicat de détecter efficacement ces domaines. Parmi ces PRMs difficilement détectables figurent les domaines Tudor (identité de séquence moyenne de l'ordre de 17%), les domaines PDZ (19%), les domaines BRCT (13%) et les domaines FHA (20%).

1.2.2 Exemple d'utilisation des PRMs pour l'intégration de signaux intra-cellulaires : le code des histones.

Le code des histones met en jeu une variété importante de modifications post-traductionnelles sur les histones telles que des phosphorylations, méthylations, acétylations, ou ubiquitinations. A ce titre, il constitue un exemple intéressant de l'utilisation des modifications post-traductionnelles spécifiquement reconnues par des domaines médiateurs d'interactions protéine-protéine dans les voies de signalisation intra-cellulaires.

Au sein de la cellule, l'ADN n'est pas nu mais compacté au sein d'une structure protéique dense appelée chromatine et composée d'ADN et d'histones, qui protège l'ADN des agressions oxydantes et régule son accessibilité aux différentes machineries cellulaires. Des modifications de la structure chromatinienne sont donc nécessaires au bon fonctionnement de tous les processus impliquant l'ADN comme la transcription, la réplication ou la réparation des dommages de l'ADN. Par exemple, suite à une cassure double brin (DSB) de l'ADN chez *Saccharomyces cerevisiae*, la phosphorylation des histones H2A au voisinage de la cassure est nécessaire à la mise en place de la réponse cellulaire adéquate (Shroff et al., 2004). Le groupement phosphate est porté par la sérine S129, située dans la région C-terminale de H2A. Lorsque ce site de phosphorylation est muté, les protéines de réparation des dommages de l'ADN ne s'accumulent pas à l'endroit de la cassure. Les trois autres histones formant le cœur de la chromatine, H2B, H3 et H4, portent elles aussi des informations capitales codées par des modifications post-traductionnelles spécifiques. La **figure 4** présente un aperçu des modifications covalentes portées par les histones et liées à la réponse aux dommages de l'ADN. Par leur présence ou leur absence, ces modifications constituent un « code des histones » reconnu par des domaines protéiques spécialisés appartenant à la

classe des PRMs qui traduisent efficacement la combinaison de ces modifications en signal cellulaire.

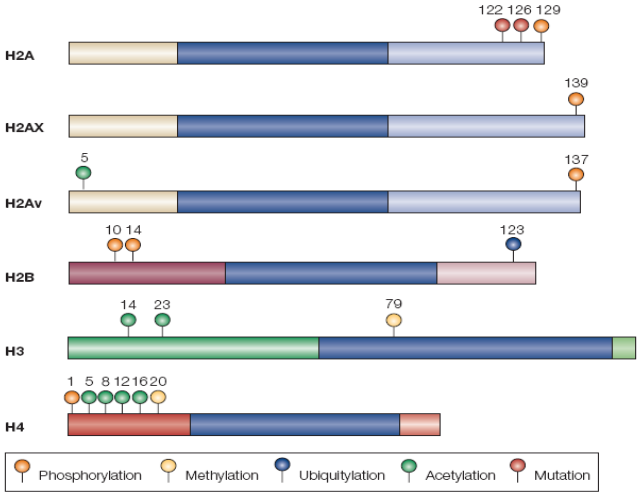


figure 4 : Modifications post-traductionnelles des histones composant leur code; d'après (van Attikum and Gasser, 2005).

Les lysines méthylées sont reconnues par des domaines Chromo, Tudor et des répétitions WD (**figure 5**-A,B,C). Ces reconnaissances sont hautement spécifiques et les domaines sont capables de discriminer les lysines méthylées en fonction de leur degré de méthylation (Kim et al., 2006; Wysocka et al., 2005). A titre d'illustration, de récentes études concernant l'interaction entre l'histone H4 et le domaine Tudor de p53BP1, ont montré que ce domaine Tudor reconnaît les lysines mono- et di-méthylées, en excluant les formes tri-méthylées (Kim et al., 2006; Wysocka et al., 2005). Les lysines acétylées des histones sont prises en charge par les domaines Bromo (**figure 5**-D) (Dhalluin et al., 1999). Enfin, alors qu'il existe une variété importante de familles de domaines capables de lier les phospho-résidus *in vivo,* seules deux d'entre elles ont été identifiées comme interagissant avec les histones phosphorylées : les domaines 14-3-3 et les tandems de domaines BRCT (**figure 5**-E,F). Cette reconnaissance est spécifique et dépend de la séquence entourant le résidu phosphorylé. En particulier, dans l'interaction entre l'isoforme 14-3-3 ζ et l'histone H3 phosphorylée sur la sérine 10, la position pS+2 est particulièrement favorable pour les prolines (Macdonald et al., 2005). De même, dans le cas du tandem de domaines BRCT de Mdc1 et de l'histone H2AX phosphorylée sur la sérine 139, le résidu en position pS+3 contribue sensiblement à l'affinité de l'interaction (Stucki et al., 2005).

D'un point de vue structural, ces études illustrent un certain nombre de caractéristiques des PRMs. L'une de celles qui nous intéresse le plus relativement à la problématique de cette thèse concerne l'origine de la spécificité de reconnaissance de ces domaines. Celle-ci est en effet remarquable. Chaque domaine est capable de reconnaître une modification covalente précise. De plus, il possède une sélectivité remarquable sur une ou deux positions supplémentaires à proximité du résidu modifié. Comprendre comment s'opère cette sélectivité afin de la prédire à partir de la structure d'un domaine constitue le thème central des chapitres 5 et 6 de ce manuscrit.

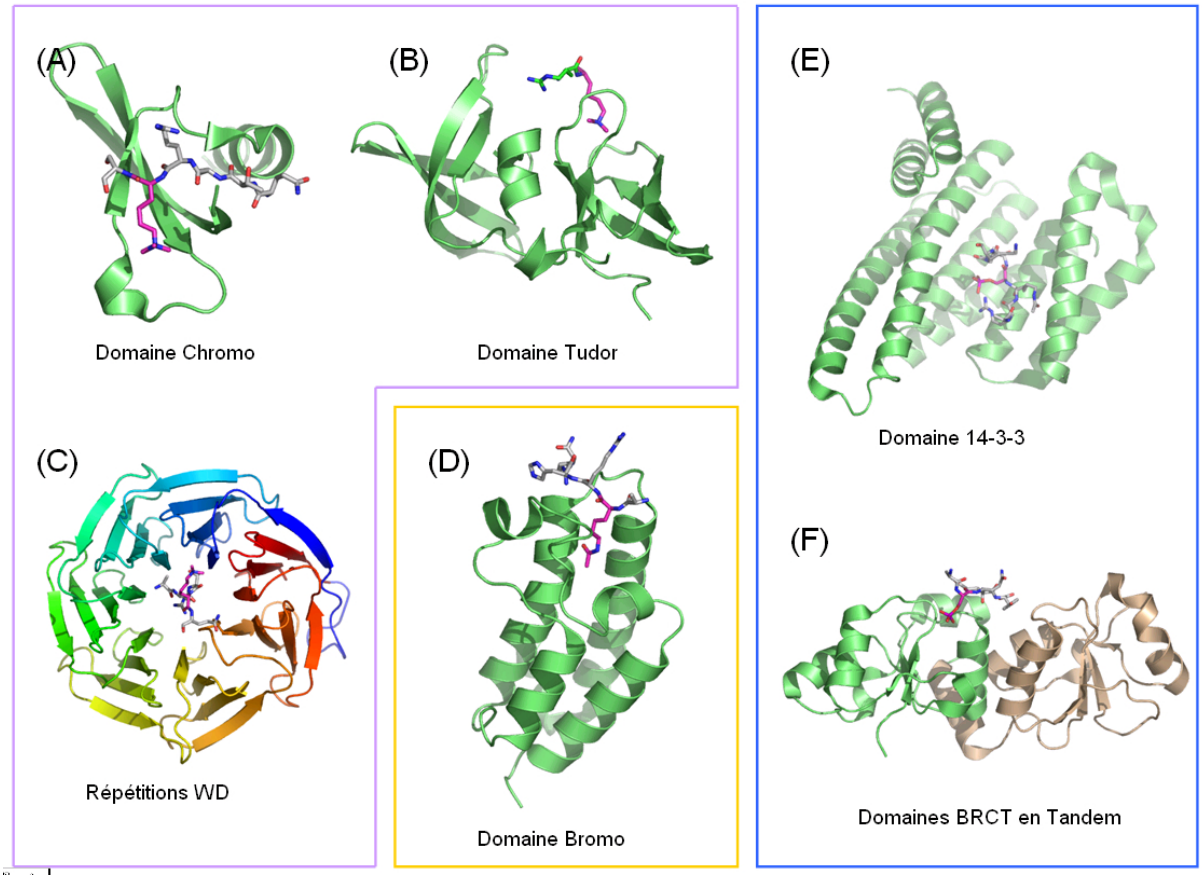


figure 5 : Structures des domaines reconnaissant des modifications post-traditionnelles des histones : les lysines méthylées (cadre violet), les lysines acétylées (cadre orange), les sérines phosphorylées (cadre bleu). Les structures A,B,C,D,E,F ont été résolues par diffraction des rayons X. Les domaines sont représentés en ruban, les fragments protéiques des histones ayant subi des modifications post-traductionnelles sont en gris, et les résidus modifiés et reconnus en roses. **(A)** Structure du domaine Chromo de la protéine de *Drosophila melanogaster* HP1 complexé à l'extrémité N-terminale de l'histone H3 tri-méthylée au niveau de la lysine 9 ; code PDB 1KNE. **(B)** Structure de la protéine humaine 53BP1 complexée avec un peptide méthylé ; code PDB 2IG0. **(C)** Structure des répétitions WD de la protéine humaine WDR5 en complexe avec un fragment de l'histone H3 méthylé sur la lysine 4 ; code PDB 2G9A, structure déposée le 6/3/2006, publication en cours de rédaction. **(D)** Structure du domaine Bromo de la protéine de levure Gcn5p complexée à un fragment de l'histone H4 acétylé sur la lysine 16 ; code PDB 1E6I. **(E)** Structure des isoformes humains 14-3-3 ζ en complexe avec un fragment phosphorylé et acétylé de l'histone H3 ; code PDB 2C1J. **(F)** Structure du tandem de domaines BRCT de la protéine humaine MDC1 (le premier domaine BRCT est en ruban vert, le second en ruban rose pâle) complexé à un fragment phosphorylé de l'histone H2AX.

1.2.3 Affinité et Spécificité des PRMs.

L'affinité et la spécificité d'une interaction entre un PRM et un fragment protéique proviennent de leur complémentarité physico-chimique, qui détermine l'énergie libre de l'interaction. Cependant, dans le contexte cellulaire, la localisation des deux interactants joue également un rôle prépondérant.

Dans le cas des PRMs dépendant de modifications post-traductionnelles, une grande partie de l'énergie de liaison doit provenir directement du groupement modifié, ce qui par définition implique que l'énergie libre de l'interaction soit faible (Bradshaw et al., 2000).

De façon plus surprenante, cette propriété est aussi partagée par les PRMs reconnaissant des motifs peptidiques sans modification post-traductionnelle. L'exemple des domaines SH3 illustre ce phénomène. Ces domaines sont spécialisés dans la reconnaissance de motifs riches en prolines et font partie des PRMs les plus étudiés à ce jour. Plusieurs équipes ont criblé des domaines SH3 par des approches de chimie combinatoire ou de *phage-display* dans le but d'identifier la séquence peptidique optimale et dans la plupart des cas, un motif consensus a pu être identifié (Cestra et al., 1999; Cheadle et al., 1994; Feng et al., 1994; Rickles et al., 1994; Sparks et al., 1994; Sparks et al., 1996). L'affinité constatée pour ces ligands spécifiques se situe dans la gamme du micromolaire, bien que dans un cas une affinité inférieure ait été constatée (Posern et al., 1998).

La construction de mutants de PRMs présentant une affinité, une sélectivité ou une spécificité différentes pour leurs substrats est récemment devenue une thématique de recherche fructueuse. En raison du volume d'informations accumulées au cours des vingt dernières années, les domaines SH3 constituent là encore un modèle d'étude de prédilection. Les travaux réalisés au sein de l'équipe de Wendell Lim (*University of California*, San Fransisco, USA) ont notamment mis en évidence que trois domaines SH3 peuvent lier des peptides synthétiques sans prolines avec une affinité et une sélectivité plus importante que celle connue pour les motifs naturels riches en prolines (Nguyen et al., 2000; Nguyen et al., 1998). En particulier, le domaine SH3 de la protéine N-Grb2 et son substrat riche en prolines naturel présentent une constante de dissociation de 5µM, tandis que pour l'un des substrats synthétiques testés cette constante atteint 0.03µM.

Cependant, une affinité de l'ordre du micromolaire ne signifie pas que l'interaction n'est pas spécifique. A titre d'exemple, les cellules humaines contiennent 14 domaines FHA, 18 domaines BRCT et 27 domaines SH3 (d'après la **table 2**). Cette abondance de PRMs d'une même famille pose la question de la spécificité de reconnaissance des peptides. Dans une publication récente, l'équipe de Wendell Lim a étudié un fragment protéique de la protéine de levure Pbs2, connu pour se lier au domaine SH3 de la protéine Sho1. Les auteurs montrent que ce fragment de Pbs2 n'interagit *in vitro* et *in vivo* avec quasiment aucun autre domaine SH3 de la levure mais peut interagir avec un certain nombre de domaines SH3 provenant de protéines d'autres organismes que la levure (Zarrinpar et al., 2003) ! Ces résultats suggèrent que la spécificité de reconnaissance des domaines SH3 pour leurs peptides est le fruit d'une combinaison de plusieurs facteurs :

- du *design* positif : au sein d'un même organisme l'affinité entre le peptide reconnu et le domaine SH3 est suffisante pour induire une interaction pour un unique domaine SH3 ;
- et du *design* négatif : l'absence de redondance entre domaines SH3 reconnaissant les mêmes peptides au sein d'un même organisme est également un critère déterminant.

Les modules spécialisés dans la reconnaissance spécifique de courts fragments protéiques sont souvent impliqués dans la régulation de l'activité des protéines qui les contiennent. Ainsi, il est vraisemblable d'inférer que les interactions PRMs-peptides sont faibles pour que ces interactions intra- et inter-moléculaires qui induisent les formes actives/inactives soient intervertibles dynamiquement (Pawson, 2004).

1.2.4 Divergence au sein des familles de PRMs.

Au sein de la cellule, le rôle des domaines médiateurs d'interactions protéine-protéine est capital. Une majorité d'entre eux est capable de reconnaître des modifications post-traductionnelles, ou plus simplement des séquences spécifiques et assurent ainsi une transmission fidèle des différents signaux intra- ou extra-cellulaires. Au regard de leur importance biologique, leur faible conservation en terme de séquence peut apparaître surprenante.

L'équipe dirigée par Rama Ranganathan (*Howard Hughes Medical Institute*, Dallas, USA) a disséqué l'espace des séquences associées aux domaines WW pour mettre en évidence les mécanismes nécessaires au repliement et au fonctionnement de ces modules connus pour se lier spécifiquement à des peptides riches en prolines (Russ et al., 2005; Socolich et al., 2005). Du point de vue structural, les domaines WW sont parmi les plus courts domaines structuraux se repliant sous la forme de monomères stables en solution sans pont disulfure ni cofacteur (environ 40 acides aminés). Ils adoptent une structure compacte sous la forme d'un feuillet de trois brins β (**figure 6**–A). La flexibilité de la boucle séparant les brins β_1 et β_2 est importante pour la formation de la poche de reconnaissance du motif poly-proline et la spécificité de la reconnaissance du motif (Peng et al., 2007). Quatre classes de domaines WW ont été répertoriées en fonction des motifs riches en prolines reconnus.

L'originalité des travaux de l'équipe de Rama Ranganathan a été d'utiliser un alignement multiple des domaines WW afin d'extraire les dépendances et corrélations statistiques entre la composition en acides aminés des différentes positions de l'alignement (**figure 6**-C). La méthode développée (SCA pour *Statistical Coupling Analysis*) leur a permis de conclure que seul un faible nombre de positions étaient sujettes à un couplage statistique et que sans surprise, les positions fortement corrélées étaient souvent voisines dans la structure (Socolich et al., 2005). Ces observations ont conduit les auteurs à poser la question suivante : la co-évolution visible en terme de séquences et ne concernant que quelques positions est-elle suffisante pour intégrer toutes les informations structurales nécessaires à un repliement de type WW ?

Pour répondre à cette question, les auteurs ont crée deux jeux de séquences artificielles de domaines WW aussi dégénérés l'un que l'autre et possédant des similitudes importantes avec des séquences de domaines WW connus (36% d'identité de séquence moyenne). Seul le second jeu de séquences artificielles respectait les contraintes imposées par l'analyse des couplages entre positions identifiés par la méthode SCA (voir **figure 6**–D et **figure 6**-E). Les résultats expérimentaux ont révélé que seules les séquences respectant les contraintes de couplage étaient repliées (un tiers des séquences du second jeu). De façon remarquable, tous les domaines repliés étaient capables de reconnaître des ligands riches en prolines appartenant aux motifs canoniques de reconnaissance des domaines WW.

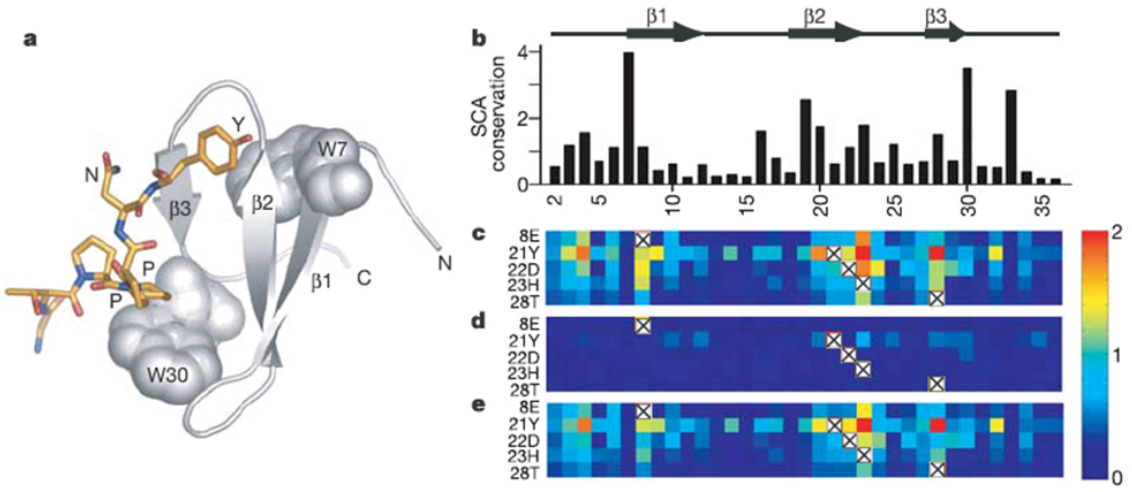


figure 6 : Etude de la relation séquence/repliement des domaines WW, par l'équipe de Rama Ranganathan. **(a)** Structure d'un domaine WW complexé à un peptide poly-proline : il s'agit du domaine WW de la protéine de rat Nedd4, code PDB 1i5h (Kanelis et al., 2001). Le domaine WW est représenté en ruban gris et le peptide cible en bâtons jaunes. Les deux tryptophanes fortement conservés au sein de la famille WW sont représentés en sphères grises. **(b)** Histogramme du score de conservation SCA de chaque position, en unité arbitraire d'énergie statistique. Le score de conservation SCA rend compte de la déviation observée entre la distribution d'acides aminés et leur occurrence moyenne dans l'ensemble des protéines (Lockless and Ranganathan, 1999). La composition en structures secondaires est indiquée au dessus de l'histogramme. **(c, d, e)** Trois représentations sous forme matricielle de la co-variation au sein de l'alignement multiple des domaines WW pour cinq positions représentées par cinq lignes (8E, 21Y, 22D, 23H, 28T) relativement à l'ensemble des 36 positions en colonne. Un gradient de couleur indique si les positions sont fortement corrélées (rouge) ou non (bleu). **(c)** Matrice dérivée de l'alignement multiple de 42 séquences de domaines WW réelles. **(d)** Matrice dérivée de l'alignement de 43 séquences artificielles de domaines WW composant le jeu 1, c'est-à-dire sans respect de la co-évolution mise en évidence par l'analyse de l'alignement multiple des WW réels. **(e)** Matrice dérivée de l'alignement multiple de 43 séquences artificielles de domaines WW composant le jeu 2, c'est-à-dire respectant les règles de co-évolution mises en évidence par l'analyse de l'alignement multiple des domaines WW réels.

Ces travaux sur les domaines WW montrent que ce n'est pas la quantité de conservation qui induit le repliement des domaines WW et sa capacité à lier les fragments poly-proline spécifiquement. A quantité de conservation égale, le respect des couplages statistiques entre différentes positions est crucial. Il est possible que cette propriété fondamentale soit applicable à d'autres domaines spécialisés dans la reconnaissance de courts fragments protéiques ; ceci permettrait d'expliquer que certains de ces domaines soient particulièrement divergents en terme de séquence tout en maintenant des stratégies d'interaction similaires (domaines Tudor, FHA, BRCT). Enfin, cette étude fournit une illustration de la versatilité évolutive qui peut se produire dans ces domaines.

1.2.5 Régulation des protéines des voies de signalisation via leurs PRMs.

Les cellules rivalisent avec les ordinateurs pour ce qui est de leur capacité à intégrer des signaux internes et externes multiples et appliquer des règles de décision complexes. De même que les circuits électroniques utilisent des composants électroniques simples, les voies de transduction des signaux cellulaires sont composées d'éléments simples comme des domaines capables d'apporter/supprimer des modifications covalentes, et d'autres domaines capables de les reconnaître. Beaucoup de protéines des voies de signalisation ont un comportement allostérique. Elles peuvent exister sous plusieurs formes, certaines actives et d'autres inactives, qui sont stabilisées par des modifications covalentes ou des ligands.

Dans le cadre des protéines des voies de transduction du signal, celles-ci contiennent généralement un domaine catalytique qui isolé montre une activité constante (non-régulée), et d'autres domaines qui inhibent l'activité de la protéine soit en bloquant l'accès au site actif du domaine catalytique (allostérie stérique), soit en contraignant la structure du domaine catalytique (allostérie conformationnelle). Ces deux types d'allostéries mettent en jeu des modules structuraux régulateurs, c'est la raison pour laquelle on les regroupe sous l'appellation « allostérie modulaire ». Pour passer d'un mode actif à un mode inactif, il suffit de relâcher l'inhibition des domaines régulateurs sur le domaine catalytique par exemple par la présence d'un ligand plus affin (**figure 7**).

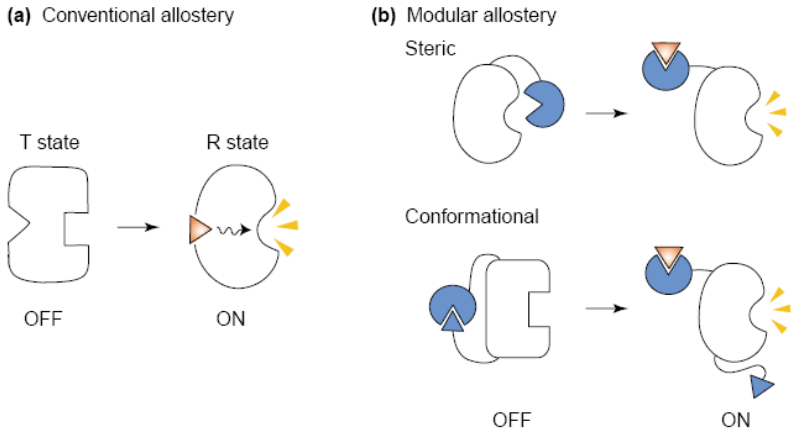


figure 7 : Comparaison entre les différents modes d'allostérie. (a) Allostérie Conventionelle. Un même domaine contient un site catalytique et un site secondaire de régulation et peut adopter une forme active ou inactive : l'amarrage du ligand au niveau du site secondaire stabilise l'une des deux formes (la forme active le plus souvent). (b) Allostérie Modulaire. Le domaine catalytique est physiquement séparé du ou des domaines régulateurs. Lorsque le domaine régulateur bloque directement l'accès au site actif afin d'auto-inhiber son activité, on parle l'allostérie modulaire stérique. Au contraire, lorsque les interactions des domaines régulateurs bloquent la conformation du site catalytique, on parle d'allostérie modulaire conformationnelle. Figure d'après (Dueber et al., 2004).

Les travaux concernant la régulation de la protéine c-Abl illustrent parfaitement la notion d'allostérie modulaire. Des travaux récents (Hantschel et al., 2003; Nagar et al., 2006; Nagar et al., 2003) ont mis en évidence les mécanismes particulièrement fins de la régulation de cette kinase composée de plusieurs domaines. La partie N-terminale regroupe un domaine SH3, un domaine SH2 et un domaine kinase ; alors que la partie C-terminale contient plusieurs domaines de liaisons. La protéine c-Abl est normalement régulée par un mécanisme d'auto-inhibition dont le dysfonctionnement peut mener à certaines leucémies.

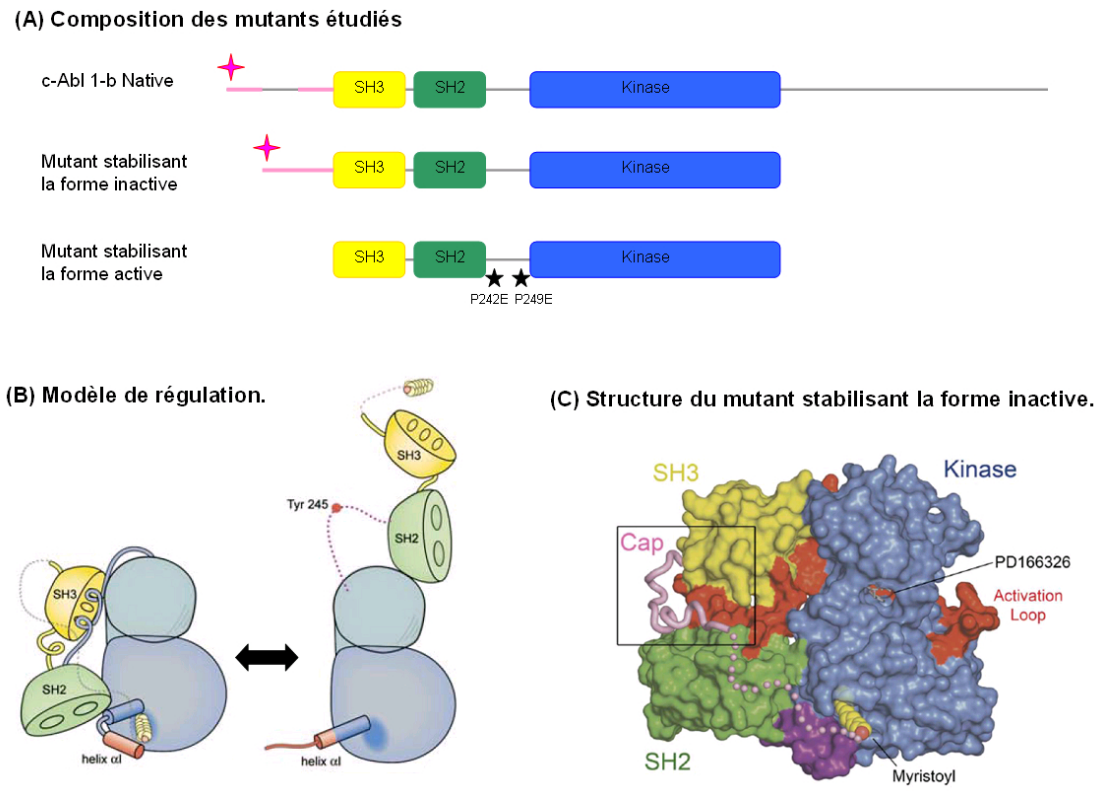


figure 8 : Régulation de la kinase C-Abl 1b par un mécanisme d'allostérie modulaire conformationnelle. **(A)** Composition de la protéine C-Abl 1b native et des deux mutants étudiés. Le domaine catalytique kinase est en bleu, les deux domaines régulateurs SH3 et SH2 sont en jaune et vert respectivement. Les sites de myristoylation sont indiqués par une croix rose, les substitutions par une étoile noire. **(B)** Modèle de régulation proposé d'après l'étude structurale et celle des différents mutants. Dans sa forme inactive (a gauche), les différentes interactions bloquent l'accès au site actif du domaine kinase (entre les deux lobes du domaine kinase, en bleu). **(C)** Structure cristallographique du mutant stabilisant la forme inactive : domaine kinase en bleu, domaine SH2 en vert, domaine SH3 en jaune, les boucles flexibles dont la boucle d'activation du domaine kinase en rouge, les hélices interagissant avec le myristol en violet. Les figures B et C sont extraites de (Nagar et al., 2006).

Grâce à deux mutants spécifiques, les auteurs ont pu étudier la structure de la protéine c-Abl stabilisée dans sa forme active et inactive (**figure 8**-A). La structure de la forme inactive de c-Abl montre que celle-ci est maintenue par trois interactions intra-moléculaires entre (i) le domaine SH2 et le domaine kinase ; (ii) le domaine SH3 et l'hélice séparant le domaine SH2

du domaine kinase ; (iii) une interaction entre une sérine phosphorylée et le domaine SH2. De plus, le complexe intra-moléculaire est rigidifié par la boucle N-terminale formée des résidus en amont du domaine SH3 et refermée par l'interaction entre le groupe myristol et le domaine SH2 (**figure 8**-B et C). L'ensemble de ces résultats a permis aux auteurs de suggérer un mode d'activation basé sur un relâchement des interactions intra-moléculaires illustré par la **figure 8**-B.

De plus en plus d'exemples de protéines des voies de signalisation dont l'activité est régulée par l'allostérie sont mis en évidence et la **table 3** récapitule les exemples pour lesquels ces mécanismes de régulation sont les mieux caractérisés à l'heure actuelle.

Protein	Input(s)	Output	Mechanism of autoinhibition
			Steric
EGFR	EGF	Receptor dimerization	Cysteine-rich domain occludes receptor dimerization surface (another cysteine-rich domain)
SH2-containing phosphatase 2 (SHP2)	SH2-binding motifs (p-Tyr)	Phosphatase	N-terminal SH2 domain sterically blocks phosphatase catalytic site
p21-activated kinase (PAK1)	Rac or Cdc42	Ser/Thr kinase	GBD blocks catalytic site, preventing autophosphorylation
Twitchin	Ca^2/S100 complex	Ser/Thr kinase	Pseudo-substrate motif occupies kinase active site; locked into position by adjacent IgG domain
p47phox	Phosphorylation by PKC	NADPH oxidase	Intramolecular peptide blocks tandem SH3 domains from interacting with membrane-associated partner, thereby blocking formation of functional oxidase complex
Vav	Phosphorylation by Src family kinases	Rho, Rac, Cdc42 GEF (DH–PH module)	N-terminal extension blocks GTPase interaction site
			Conformational
Src kinases	SH2- and SH3-binding motifs	Tyr kinase	Binding of the SH2 and SH3 domains to intramolecular ligands locks kinase in inactive conformation
c-Abl	SH2- and SH3-binding motifs; possibly membrane targeting of myristoyl group	Tyr kinase	Binding of N-terminal myristoyl group and SH2 and SH3 domains to sites on or adjacent to kinase domain locks kinase in inactive conformation remarkably similar to the autoinhibited structure of Src
N-WASP	Cdc42 and PIP2	Arp2/3 stimulation (actin polymerization)	GBD and a polybasic motif (B) form cooperative intracomplex interactions that conformationally inactivate the N-WASP output domain, blocking its ability to activate the Arp2/3 actin-nucleating complex

table 3 : Exemples de protéines dont l'action est régulée par auto-inhibition. La partie haute rassemble les exemples où la régulation se fait par allostérie modulaire stérique. La partie basse comporte les exemples où la régulation se fait par allostérie modulaire conformationnelle, comme c-Abl.

La métaphore entre protéines régulées et « portes logiques » est fréquemment utilisée : de même qu'une protéine intègre plusieurs signaux et les combine de manière à en déduire quel doit être son état d'activité (actif ou inactif), une porte logique intègre elle aussi plusieurs signaux qu'elle analyse de manière à produire un signal de sortie. Par exemple, la **figure 9**-A illustre comment la protéine N-WASP, qui dans son état actif est capable d'activer la polymérisation de l'actine *via* le complexe Arp2/3, peut être comparée à une porte logique « AND ». L'activité de N-WASP est régulée par deux domaines régulateurs : un domaine GDB et un domaine basique. C'est uniquement si les substrats spécifiquement reconnus par les deux domaines GDB et B sont présents simultanément que l'activité de N-WASP est maximale. A l'inverse, dès qu'un des deux substrats est absent, l'activité est très réduite (Prehoda and Lim, 2002). Cette propriété intégrative surprenante est d'autant plus importante que l'on considère des concentrations de substrats faibles : si la concentration en substrats en très inférieure au Kd alors le fonctionnement intégratif est conséquent, tandis qu'il est quasiment nul lorsque la concentration est supérieure au Kd.

L'utilisation de l'allostérie modulaire semble particulièrement adaptée d'un point de vue évolutif car elle permet de produire des « portes logiques » facilement « reprogrammables ». En particulier, l'équipe dirigée par Wendell Lim (*University of California*, San Fransisco, USA) a testé s'il était possible de reprogrammer le comportement d'une protéine régulée (Dueber et al., 2003). Les auteurs ont synthétisé des protéines hydrides comprenant (**figure 9**-B):

(i) le domaine actif de la protéine N-WASP ;
(ii) deux PRMs (SH3 et PDZ) ;
(iii) les fragments reconnus par ces deux modules.

L'ordre dans lequel ces différents éléments sont inclus dans les protéines hybrides varie, de même que la longueur des linkers séparant les différents domaines. Les auteurs mettent en évidence que 2/3 des protéines conçues *via* ce système ont un comportement de « porte logique », c'est-à-dire que leur activité catalytique varie en fonction de l'ajout des ligands des deux PRMs. Les auteurs observent également que ce système permet de mimer des portes logiques simples (« AND », « OR »), mais aussi des portes plus sophistiquées, comme la porte « NAND NOT ».

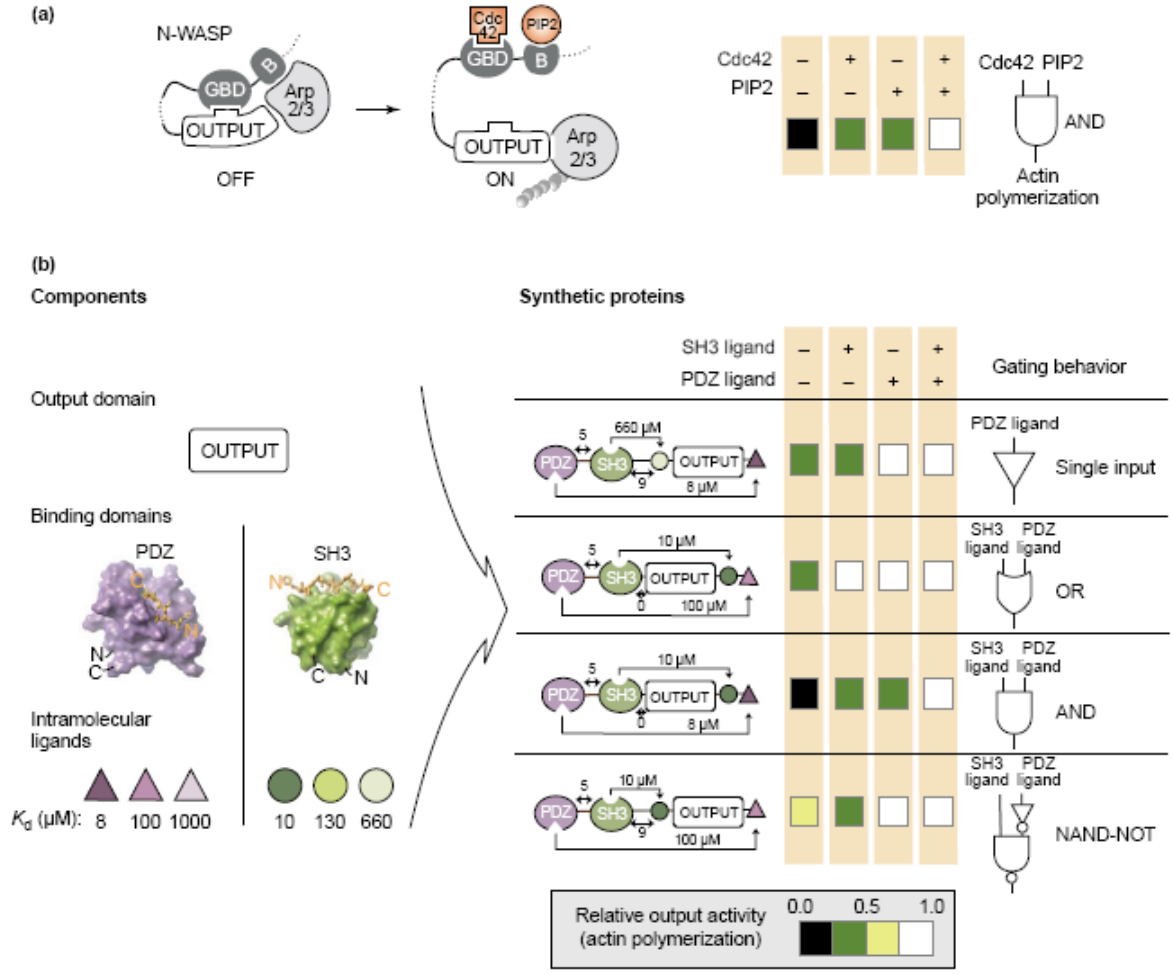


figure 9 : Jeu de « portes logiques » à partir de la protéine N-WASP. **(a)** La protéine N-WASP possède deux sous-unités régulatrices notées GDB et B. Lorsque ces deux domaines forment des interactions intra-moléculaires avec le domaine catalytique et la protéine Arp2/3, l'activité du domaine catalytique est nulle. En présence de l'un ou l'autre des substrats extra-moléculaires de GDB et B, l'activité est très faible. Par contre, lorsque les deux substrats extra-moléculaires sont présents simultanément, les domaines GDB et B relâchent leurs interactions intra-moléculaires pour se lier aux substrats extra-moléculaires plus affins, et par ce fait le domaine catalytique devient complètement actif. On mime donc une porte logique « AND » puisque l'activité n'est complète qu'en présence simultanée des deux substrats extra-moléculaires. **(b)** Le domaine catalytique de N-WASP est associé à deux PRMs, les domaines PDZ et SH3. Des peptides se liant aux domaines PDZ et SH3 sont insérés dans la séquence afin de permettre des interactions intra-moléculaires. Dans la première construction, l'activité catalytique de la protéine ne dépend que de l'ajout du substrat du domaine PDZ. Dans la seconde construction, le domaine catalytique est actif dès que l'un des deux substrats extra-moléculaires est ajouté. La troisième construction est quasiment identique à la deuxième : les différents éléments se succèdent dans le même ordre mais les linkers sont plus courts. Le domaine catalytique est actif uniquement lorsque les deux substrats sont ajoutés simultanément. Enfin, la quatrième construction est active lorsque le substrat du domaine PDZ est absent, ou lorsque les deux substrats sont absents simultanément. Figure d'après (Dueber et al., 2004).

D'autres approches ont été testées : (i) insérer de nouveaux domaines au sein d'une protéine dont l'activité est déjà régulée par des mécanismes d'allostérie modulaire (Guntas and Ostermeier, 2004; Radley et al., 2003), (ii) créer des protéines hybrides soit en utilisant des mécanismes simples de recombinaisons non-homologues soit en fusionnant des séquences (Guntas et al., 2005; Guntas et al., 2004; Sallee et al., 2007). L'ensemble de ces résultats récents concordent avec ceux de l'équipe de Lim et montrent qu'il est possible de reprogrammer le comportement d'une protéine « porte logique » en modifiant sa composition en domaines et leur arrangement. L'hypothèse de Lim est que cette facilité à modifier le type de « porte logique » en jouant sur la recombinaison des différents PRMs et de leurs substrats est une clé des mécanismes d'évolution des systèmes de signalisation (Dueber et al., 2004).

1.3 Développements bioinformatiques pour prédire les propriétés des PRMs : Objectif de la thèse.

Nous avons souligné dans le chapitre précédent que les modules spécialisés dans la reconnaissance spécifique de peptides jouent un rôle primordial dans la régulation intra- et inter-moléculaire des protéines des voies de signalisation. En plus de ce rôle biologique crucial, ils se caractérisent par des propriétés telles que :

- une divergence importante voire très importante au sein des séquences d'une même famille, malgré un repliement conservé ;
- une énergie d'interaction domaine/ peptide faible, malgré une spécificité de reconnaissance remarquable ;
- une capacité à combiner leurs actions pour intégrer différents signaux et, par des mécanismes d'allostérie conformationnelle complexes, réguler l'activité catalytique des protéines selon des règles de décision subtiles.

Ces différents aspects rendent l'étude des PRMs particulièrement intéressante. Le travail présenté dans cette thèse s'intéresse plus particulièrement aux approches bioinformatiques permettant de guider l'étude expérimentale des PRMs.

Dans une première partie de cette thèse, nous nous interesserons au problème de la détection et de la modélisation structurale des PRMs. Nous avons souligné que certaines familles de PRMs étaient très divergentes en termes de séquence ; or les processus actuels de détection et modélisation de domaines structuraux reposent sur des méthodes d'alignement de séquences qui ne sont pas fiables lorsque la conservation des séquences au sein de la famille est faible (la limite inférieure de 25% d'identité de séquence est souvent citée). Je me suis intéressée à cette question et aux moyens de dépasser les limitations actuelles des méthodes d'alignement (chapitres 2 et 3). Pour aborder cette problématique, les concepts fondamentaux des méthodes d'alignement seront rappelés dans la suite de l'introduction.

Le chapitre suivant (chapitre 4) sera consacré à la recherche des sites d'interaction des PRMs sur leurs partenaires. Pour cela, nous nous sommes focalisés sur l'étude de la protéine Rad53, kinase essentielle des voies de surveillance des dommages de l'ADN chez la levure *Saccharomyces cerevisiae*. Cette protéine est régulée par deux domaines FHA qui encadrent le

domaine catalytique central. Rad53 se situe au cœur d'un réseau d'interactions protéine-protéine dense et interagit avec plus d'une dizaine de partenaires par l'intermédiaire de ses domaines FHA. Nous avons développé une stratégie bioinformatique automatique visant à prédire les sites reconnus par les domaines FHA de Rad53 sur chacun de ses partenaires. Pour que ces prédictions puissent être validées expérimentalement, plusieurs collaborations internes à l'iBiTecS ont été initiées avec l'équipe de Marie-Claude Marsolier-Kergoat.

Enfin, les deux derniers chapitres (chapitres 5 et 6) abordent le problème délicat de la prédiction automatique et semi-automatique des motifs spécifiquement reconnus par les PRMs. Les techniques actuelles utilisées pour prédire le motif le plus affin pour un PRM donné sont basées sur des approches expérimentales comme le criblage de bibliothèque de peptides (méthodes *in vitro*) ou le phage display (méthode *in vivo*). Ces deux techniques présentent des désavantages car les méthodes *in vitro* sont lentes et coûteuses à mettre en place alors que les méthodes *in vivo* ne permettent pas de traiter le cas des PRMs reconnaissant des modifications post-traductionnelles. Le bénéfice d'une méthode bioinformatique permettant de prédire la spécificité d'un PRM est donc certain. La prédiction d'une spécificité d'interaction entre un domaine et son ligand s'apparente à une problématique de *design* pour laquelle on cherche à maximiser la stabilité d'un complexe protéine /ligand par des mutations dans la séquence du ligand. Des approches basées sur les algorithmes de *design* automatique ou semi-automatique des interfaces PRM/peptide ont été développées et ont donné des résultats intéressants lorsque la région du PRM située à l'interface du peptide est une région de structures secondaires rigides (Wiedemann et al., 2004). Nous nous sommes intéressés au cas plus général des PRMs dans lesquels la région à l'interface du peptide est une région flexible, comme c'est notamment le cas des domaines FHA et BRCT. Cet aspect n'ayant pas été abordé précédemment en raison de sa complexité, cette étude nous permettra de tester les limites actuelles des méthodes de *design* et de proposer de nouvelles stratégies de prédiction pour la bioinformatique structurale.

Un certains nombre d'outils bioinformatiques ont été développés ces dernières années pour traiter des problèmes récurrents que sont la détection de domaines protéiques et leur modélisation ou le *design* de structures. Dans la suite de cette introduction, les méthodes bioinformatiques actuellement utilisées pour la détection et la modélisation de domaines protéiques sont présentées dans le paragraphe 1.4. Les paragraphes 1.5 et 1.6 synthétisent l'état des méthodes existantes pour résoudre deux problèmes sous-jacents au *design* :

comment générer la structure la plus stable possible en optimisant le placement des chaînes latérales et comment approximer le plus précisément possible l'énergie libre associée aux modèles produits.

1.4 Méthodes visant à prédire le repliement associé à une séquence.

1.4.1 Introduction.

Plus de 35 000 structures protéiques résolues expérimentalement sont actuellement référencées dans la PROTEIN DATA BANK (Berman et al., 2000). Cependant et malgré leur nombre en constante augmentation, le fossé entre le nombre de structures résolues et le nombre de séquences protéiques disponibles ne cesse de croître. Pour cette raison, la prédiction de la structure tridimensionnelle des protéines par des approches bioinformatiques est devenue un sujet de recherche majeur depuis une dizaine d'années.

Le premier test à grande échelle permettant de comparer les différentes méthodes de prédiction de la structure tridimensionnelle des protéines a été organisé en 1994 à l'initiative de John Moult, Michael James, Shoshana Wodak et Fred Cohen. Cette expérience a pris le nom de CASP, signifiant *Critical Assesment of protein Structure Prediction*. Son déroulement est organisé en 3 phases :

(i) collecte des cibles, c'est-à-dire des structures récemment résolues et n'ayant pas encore été publiées ;

(ii) prédiction en un temps limité de la structure tridimensionnelle des protéines cibles collectées par les différentes équipes travaillant sur la prédiction, sans aucune indication sur les structures expérimentalement résolues ;

(iii) évaluation de la qualité des prédictions faites par chacun des groupes pour chacune des cibles, en comparant les prédictions et les structures expérimentales.

Cette expérience a permis pour la première fois de comparer les différentes approches développées et s'est avérée très constructive. Depuis 1994, une nouvelle session de CASP est organisée tous les deux ans (Moult, 2005; Moult et al., 2005; Moult et al., 2001; Moult et al., 2003; Moult et al., 1997; Moult et al., 1999), et une session nommée CAFASP a été introduite

pour évaluer les méthodes de prédictions entièrement automatisées (Fischer et al., 1999). Ceci a l'avantage d'introduire de l'émulation dans la communauté des bioinformaticiens structuralistes et de mettre en avant les progrès réalisés au cours des dernières années. Dans la suite, nous introduisons les principales approches utilisées pour la prédiction de la structure tridimensionnelle des protéines.

1.4.2 La modélisation comparative.

La modélisation comparative (**figure 10**) utilise une relation évolutive entre la séquence que l'on souhaite modéliser, dite « séquence cible », et une autre séquence dont la structure est connue, dite « séquence de référence ». Cette relation évolutive, traduite via un alignement des deux séquences, permet de modéliser la structure associée à la " séquence cible " en évaluant l'impact sur la structure de référence des mutations, insertions et délétions qui différencient la « séquence cible » de la « séquence de référence ».

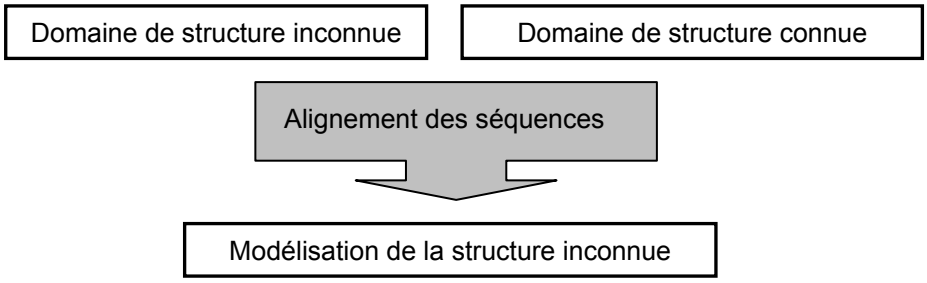


figure 10 : Principe général de la modélisation comparative. La modélisation du repliement associé à la " séquence cible " est déduite de la séquence et de la structure d'un ou plusieurs de ses homologues, dites " séquence(s) de référence " et " structure(s) de référence ".

La procédure de modélisation comparative se divise en 4 étapes que nous allons décrire brièvement.

Etape 1 : détection et choix des structures de référence. Puisque la structure associée à la séquence cible est modélisée à partir d'une ou plusieurs structures de référence, il est très important de choisir judicieusement la ou les structure(s) de référence. Depuis quelques années, des progrès importants ont été effectués dans ce domaine notamment grâce aux méthodes de comparaison profil-profil (Debe et al., 2006; Ginalski and Rychlewski, 2003;

Ginalski et al., 2004; Pietrokovski, 1996; Rychlewski et al., 2000; Sadreyev and Grishin, 2003; Yona and Levitt, 2002) et HMM-HMM (Soding, 2005). L'utilisation de ces méthodes a notamment permis, lors de l'édition de CASP 2003, de repérer une relation d'homologie très lointaine entre la séquence à modéliser et une séquence de structure connue, alors que l'identité de séquence était de 6% (Tramontano and Morea, 2003) ! On constate au travers de cet exemple que si une structure homologue même lointaine existe, une combinaison judicieuse des méthodes actuelles permet de la détecter.

Etape 2 : alignement des séquences. L'alignement optimal du point de vue de la procédure de modélisation comparative est celui qui serait déduit de la superposition des structures de la cible et des structures de référence. Bien entendu, on ne dispose pas de cet alignement puisque la structure de la cible n'est pas connue.

Différentes méthodes ont été développées afin de permettre d'aligner au mieux les séquences sans connaître leur structure. On note qu'à l'heure actuelle, on ne dispose d'un alignement de qualité que lorsque la séquence cible et les séquences de référence partagent un fort taux d'identité de séquence (supérieur à 30%). En dessous de cette limite, les alignements générés ne sont pas fiables et par conséquent les modèles produits ne sont pas de bonne qualité. Nous détaillerons dans les paragraphes suivants les différentes méthodes d'alignement de séquence.

Etape 3 : construction du modèle. De nombreuses méthodes ont été développées afin de modéliser au mieux la structure d'une séquence cible une fois les structures de référence choisies et l'alignement des séquences effectué (Sali and Blundell, 1993; Schwede et al., 2003). Parmi celles-ci, le programme le plus utilisé et le plus performant est MODELLER (Sali and Blundell, 1993).

Mis au point par l'équipe d'Andrej Sali en 1993, MODELLER extrait un ensemble de contraintes spatiales à partir des structures de référence et de l'alignement des séquences, auxquelles s'ajoutent des contraintes stéréochimiques (longueurs et angles de liaison) et statistiques (préférences pour certaines distances inter-atomiques non-liées). La structure associée à la séquence cible est alors construite de manière à respecter au mieux ces contraintes spatiales (**figure 11**). Plus précisément, les contraintes spatiales, exprimées sous

la forme de fonctions de densité de probabilité, sont combinées au sein d'une fonction objective qui est optimisée par un recuit simulé associé à des simulations de dynamique moléculaire.

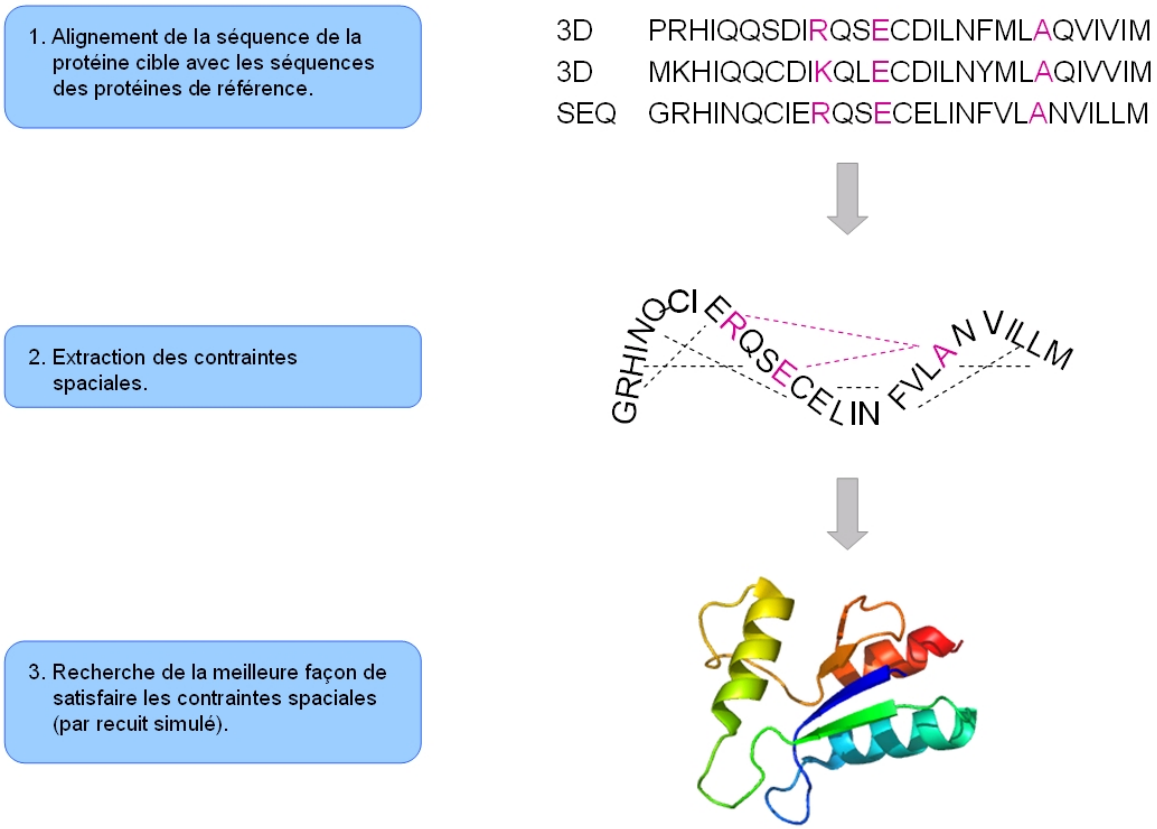


figure 11 : Illustration de la méthode permettant à MODELLER de construire un modèle du squelette peptidique d'une protéine cible " SEQ " à partir de deux structures de références "3D" , d'après (http://salilab.org/modeller/manual/node13.html). (1) La séquence protéique cible est alignée aux séquences protéiques des structures de référence. (2) Les contraintes spatiales appliquées à la séquence cible sont déduites des structures de référence et de l'alignement obtenu à l'étape 1. (3) On détermine la structure de la séquence cible de façon à satisfaire au mieux les contraintes spatiales précédemment déterminées.

Etape 4 : optimisation du modèle. Les problèmes majeurs de la modélisation comparative proviennent d'erreurs d'alignement. Cependant, même lorsque l'alignement utilisé est optimal, le modèle construit comporte souvent des imprécisions qu'il est important de corriger par une étape d'optimisation. La modélisation des boucles séparant deux éléments de structure secondaire consécutifs pose notamment problème et nécessite l'utilisation de méthodes dédiées (Fiser et al., 2000; Fiser and Sali, 2003; van Vlijmen and Karplus, 1997). Il est également possible de raffiner le modèle produit en effectuant une étape de prédiction de la conformation des chaînes latérales (Canutescu et al., 2003; Dunbrack and Karplus, 1993; Mendes et al., 1999; Mendes et al., 2001; Xiang and Honig, 2001) où en soumettant la

structure modélisée à une étape de minimisation dans un champ de force existant (Brooks et al., 1983; Van Der Spoel et al., 2005).

1.4.3 Alignement de séquences et modélisation comparative.

Dans le contexte de la modélisation comparative, l'alignement de séquence est une étape limitante. Les résultats du concours CASP en témoignent : beaucoup d'équipes raffinent encore manuellement les alignement qu'elles utilisent (Tramontano and Morea, 2003).

Pour aligner une séquence s_{obs}, de structure inconnue sur la séquence t_{ref} d'une structure connue et réaliser un modèle de s_{obs}, le meilleur alignement possible entre s_{obs} et t_{ref} au vu de la procédure de modélisation comparative serait l'alignement issu de la superposition des structures de s_{obs} et t_{ref}. Cependant, cet alignement structural n'est pas connu puisque la structure de s_{obs} n'est pas déterminée. On cherche donc à l'approximer à l'aide d'alignements de séquence dont cette section reprend l'historique.

1.4.4 Les alignements de séquence à séquence, ou alignements par paires.

Si le problème se restreignait à déterminer l'alignement, sans insertions ni délétions, tel que le nombre de paires de résidus identiques soit maximal, il pourrait être résolu efficacement par l'algorithme de Bellman (1957) qui permet par programmation dynamique de trouver le plus court chemin depuis un sommet source dans un graphe orienté pondéré (Bellman, 1957). Cependant, un alignement de séquences protéiques pertinent ne se résume pas à une mise en correspondance optimale du point de vue de l'identité de deux séquences : il est également nécessaire de tenir compte de la « similitude » entre acides aminés, et de trouver un traitement adéquat pour les zones d'insertions et de délétions.

Les matrices de substitution. Elles permettent de réaliser une pondération des remplacements d'un acide aminé par un autre et contiennent donc 20*20=400 scores. Plusieurs matrices ont été proposées. Certaines sont basées sur les caractéristiques physico-chimiques des acides aminés (Barker et al., 1990; Grantham, 1974; Miyata et al., 1979; Mohana Rao, 1987), tandis que d'autres ont été construites à partir de substitutions observées

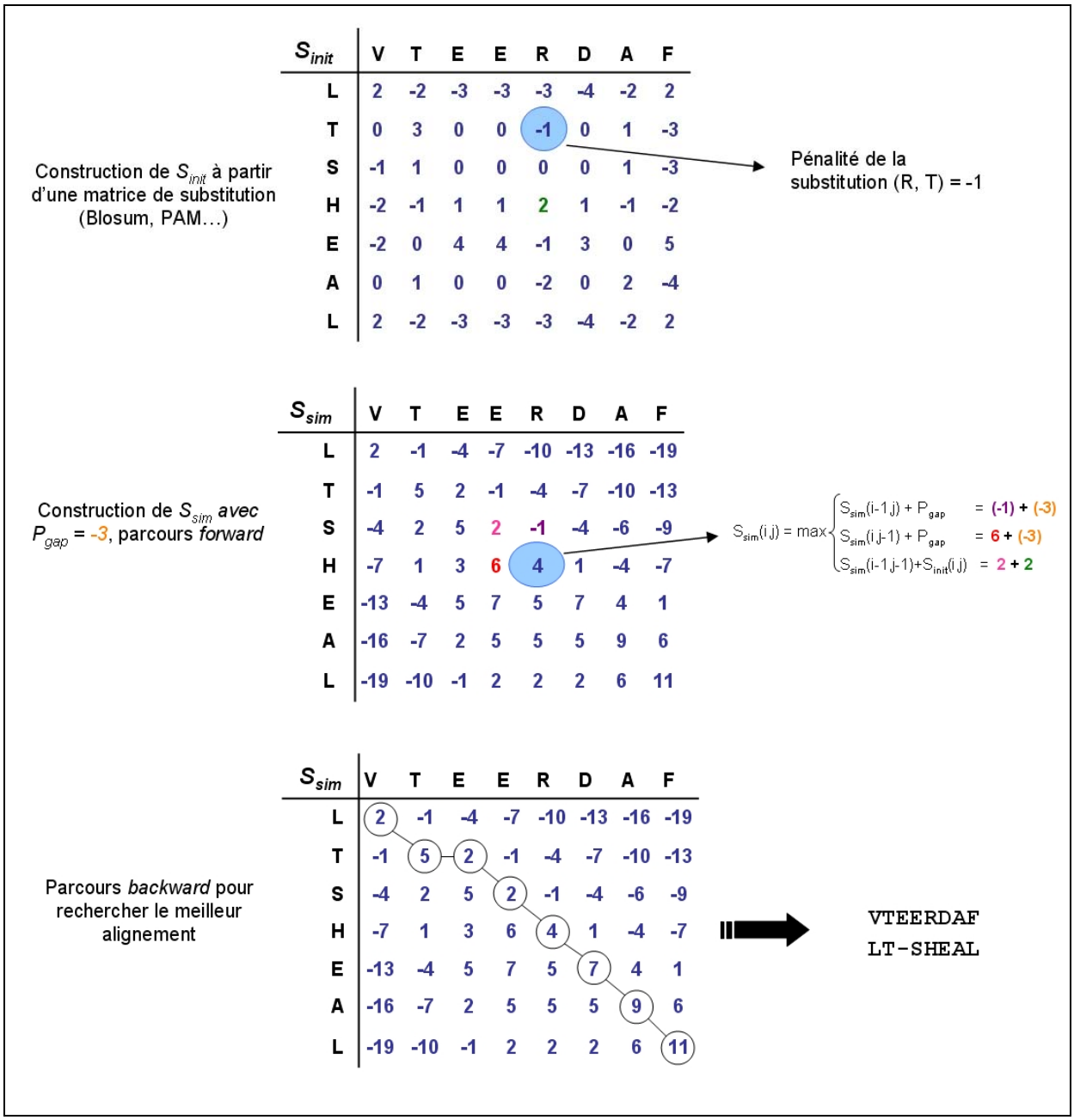
au sein d'un ensemble de familles de séquences alignées (Dayhoff, 1973; Dayhoff, 1978; Henikoff and Henikoff, 1992; Kosiol and Goldman, 2005) ou de structures superposées (Overington et al., 1990; Risler et al., 1988). Le consensus général est que les matrices dérivées des données de substitution sont plus efficaces que celles fondées sur d'autres critères (Henikoff and Henikoff, 1993).

Traitement des insertions et délétions par l'utilisation d'une fonction affine. La valeur attribuée aux pénalités d'insertions est généralement calculée par une fonction affine de type $uk+v$; où v représente la pénalité d'ouverture de l'insertion et u la pénalité associée à la prolongation d'une insertion existante. En choisissant $v>>u$; les alignements privilégient un modèle avec peu de régions d'insertions, celles-ci pouvant être longues. Cette méthode de prise en compte des insertions/délétions a l'avantage d'être simple et facile à mettre en œuvre, c'est la raison pour laquelle elle est majoritairement utilisée. Néanmoins, dans le cadre des alignements de séquences très divergentes, la modélisation adéquate des insertions reste un domaine de recherche toujours actif (Chang and Benner, 2004; Nozaki and Bellgard, 2005; Wrabl and Grishin, 2004; Zachariah et al., 2005).

Algorithme de Needleman et Wunsch, algorithme de Sellers. Needleman et Wunsch (1970) puis Sellers (1974) ont proposé les premiers algorithmes d'alignement de deux séquences protéiques dans lesquels furent prises en compte les similarités entre acides aminés ainsi que le traitement des insertions et délétions par une fonction affine (Needleman and Wunsch, 1970; Sellers, 1974; Sellers, 1974). Les deux algorithmes sont très proches : le premier maximise la similitude entre les deux séquences alors que le second minimise leur distance. En 1981, l'équivalence de ces deux algorithmes pour résoudre le problème de l'alignement de séquences protéiques optimal a été démontrée (Smith et al., 1981).

Dans sa forme actuelle, l'algorithme de Needleman et Wunsch est basé sur l'utilisation d'une matrice de substitution mesurant la similarité entre résidus deux à deux couplée à une fonction affine $uk+v$ pour rendre compte des insertions et délétions. A l'aide de cela, une matrice de similarité S est construite selon le principe explicité dans **figure 12** (Needleman and Wunsch, 1970). Une fois la matrice de similarité S_{sim} construite (étape *forward*), un parcours en sens inverse de S_{sim} permet de déterminer le meilleur alignement (étape *backward*, ou de *backtracking*).

Par la suite, d'autres algorithmes directement inspirés de celui de Needleman et Wunsch ont été introduits, par exemple pour identifier le meilleur alignement local entre deux séquences protéiques (Smith and Waterman, 1981). Dans une étude publiée en 1983, Walter Fitch et Temple Smith notent que ces méthodes d'alignements par paires dépendent crucialement des valeurs de la matrice de substitution et de la valeur du couple *(u,v)* appliquée pour le traitement des insertions et délétions (Fitch and Smith, 1983).



S_{init}	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

S_{sim}	V	T	E	E	R	D	A	F
L	2	-1	-4	-7	-10	-13	-16	-19
T	-1	5	2	-1	-4	-7	-10	-13
S	-4	2	5	2	-1	-4	-6	-9
H	-7	1	3	6	4	1	-4	-7
E	-13	-4	5	7	5	7	4	1
A	-16	-7	2	5	5	5	9	6
L	-19	-10	-1	2	2	2	6	11

figure 12 : Algorithme de Needleman et Wunsch illustré pour trouver l'alignement optimal global entre deux séquences protéiques. Dans une première étape, dite étape *forward*, la matrice de similarité est calculée : les scores sont calculés de gauche à droite et de haut en bas selon la formule explicitée à droite. Ensuite, dans un second temps, à l'étape *backward*, le parcours optimal est identifié est commencant en bas à droite et en remontant dans le matrice de similarité.

1.4.5 Les alignements de séquences à séquences alternatifs et sous-optimaux.

Dans le contexte de la modélisation comparative, l'alignement idéal est l'alignement structural. Lorsque la structure des deux partenaires n'est pas connue, on cherche à approximer l'alignement structural par l'alignement des séquences d'après l'algorithme présenté précédemment. Mais lorsque les séquences à aligner sont très divergentes, il est fréquent que l'alignement optimal du point de vue des séquences ne coïncide pas avec l'alignement structural.

Pour chercher des alignements plus proches de l'alignement structural dans le voisinage de l'alignement des séquences optimal, des algorithmes cherchant à générer des alignements alternatifs entre deux séquences ont été développés.

Méthode paramétrique. La méthode la plus évidente pour générer plusieurs alignements au lieu d'un seul est de faire varier les paramètres comme la matrice de substitution, ou la valeur du couple *(u,v).* (Fitch and Smith, 1983; Gusfield et al., 1992; Waterman et al., 1992). En effet, puisqu'il a été montré que les méthodes d'alignements *pairwise* dépendent fortement de ces paramètres (Fitch and Smith, 1983), les modifier permet de perturber les alignements produits.

Algorithme de Saqi et Sternberg. Saqi et Sternberg ont proposé en 1991 une heuristique itérative nommée *Iterative Elimination Method* et permettant de générer des alignements alternatifs proches de l'alignement optimal (Saqi and Sternberg, 1991). L'idée est la suivante : après avoir identifié le meilleur alignement à l'aide de l'algorithme de Needleman et Wunsch, la matrice de similarité S_{sim} est modifiée afin de pénaliser d'un facteur Δ toutes les cellules par lesquelles passe le meilleur alignement. Un nouveau parcours *backward* de S_{sim} permet alors d'identifier le meilleur alignement sur la matrice modifiée. Si celui-ci est identique au précédent, cela témoigne d'une certaine robustesse de l'alignement optimal ; tandis que si celui-ci est différent, alors il constitue un alignement alternatif. Pour que l'exploration soit plus importante, les auteurs utilisent cette méthode en faisant varier le couple *(u,v)* ainsi que le paramètre Δ.

Algorithme de Zücker. En 1991 également, Mickaël Züker (*Institute for Biological Sciences*, Ottawa, Canada) a proposé un autre algorithme ingénieux pour générer des alignements alternatifs dont les scores mathématiques soient néanmoins élevés (Zuker, 1991). Soient s_{obs} = s_1 ..s_M et t_{ref} = t_1..t_N les deux séquences à aligner, l'idée est la suivante : une matrice *S1* stocke le résultat du parcours *forward* de *(s_1..s_M, t_1..t_N)*, tandis qu'une matrice *S2* stocke le résultat du parcours *forward* des deux séquences renversées *(s_M..s_1, t_N..t_1)*. Ainsi, pour un couple *(i,j)* donné, on connaît à la fois le meilleur chemin allant jusqu'à *(i,j)* stocké dans la matrice *S1*, et le meilleur chemin repartant de *(i,j)* stocké dans la matrice *S2*. En combinant les matrices *S1* et *S2*, il est donc possible de déterminer pour tout couple *(i,j)*, le meilleur alignement entre s_{obs} et t_{ref} passant par *(i,j)*. En imposant différents couples *(i,j)* ou en les explorant tous si la longueur des deux séquences n'est pas excessive, il est possible de générer un grand nombre d'alignements alternatifs.

Algorithme de Waterman et Byers. L'algorithme développé par Michaël Waterman (*University of South Califormia*, Los Angeles, USA) et Thomas Byers (*Digital Research Inc*. Pacific Grove, USA) est une variante de l'algorithme de Sellers permettant de déterminer l'ensemble des alignements dont le score est proche du meilleur score. Ainsi, puisque l'algorithme de Sellers minimise la distance d séparant s_{obs} et t_{ref} ; l'algorithme de Waterman et Byers identifie l'ensemble $\mathcal{E}$ des alignements dont le score est compris entre la distance minimale d_0 et $d_0+\varepsilon$. Seule une modification de la règle de décision locale permettant de « remonter » au sein de la matrice de distance lors de la procédure de *backtracking* est nécessaire, mais il convient de porter une attention particulière aux structures de données et à la mémoire requise pour stocker les différents parcours (Waterman and Byers, 1985).

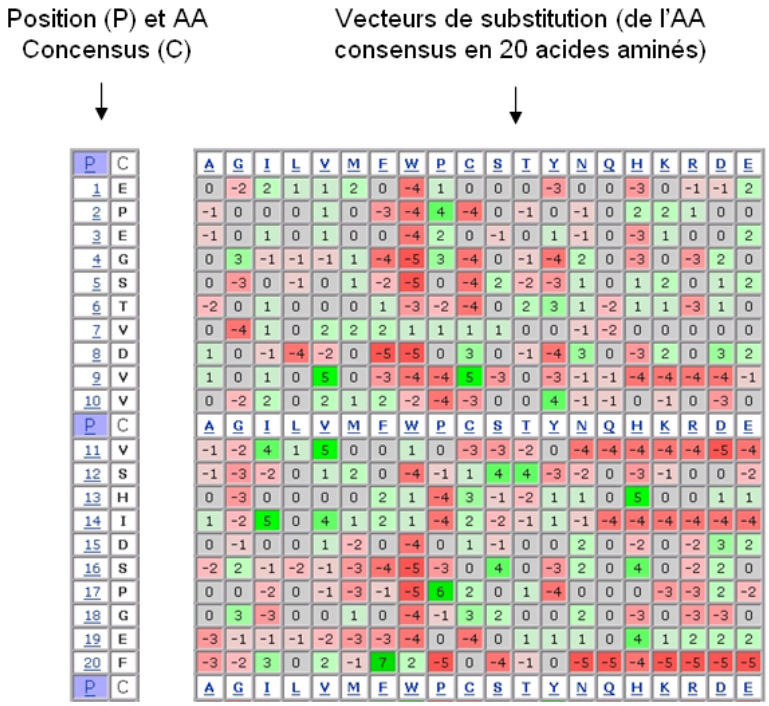
Les alignements *pairwises* sous optimaux ont été étudiés et plusieurs approches se sont distinguées. L'algorithme de Waterman est Byers est le seul à fournir le voisinage optimal exact de l'alignement de séquence optimal : soit $\mathcal{E}$ l'ensemble d'alignements générés, il atteste qu'il n'existe aucun alignement qui ne fasse pas partie de $\mathcal{E}$ et qui ait un score plus élevé qu'un alignement de $\mathcal{E}$. Les trois autres approches présentées ne garantissent pas cette propriété, ce qui induit le risque de « passer à côté » de l'alignement que l'on recherche en explorant les alignements sous optimaux. Cependant, l'intérêt des algorithmes de Züker, de Saqi et Sternberg et surtout de l'approche paramétrique est de générer une plus grande diversité d'alignements.

1.4.6 Les alignements d'une séquence sur un alignement multiple de séquences : séquence-profil, séquence-HMM.

Les domaines protéiques partageant un même repliement possèdent une empreinte de ce repliement le long de leur séquences : certaines positions clefs sont contraintes car leur substitution affecterait le repliement, tandis que d'autres sont plus variables. Les alignements de séquences par paires ne permettent pas de refléter cette propriété puisque quelle que soit la position le long de l'alignement, une même substitution a toujours le même score. Les approches basées sur les alignements multiples permettent de corriger cela et d'accorder plus d'importance aux régions conservées au sein d'une famille de domaines. Parmi les approches modélisant les alignements multiples, on distingue les méthodes utilisant des matrices PSSM (pour *Position Specific Scoring Matrix*), et celles basées sur les modèles de Markov cachés (ou HMM pour *Hidden Markov Model*)

Les matrices PSSM, ou profils. Intuitivement, on peut identifier une matrice PSSM à un ensemble de vecteurs de substitution spécifiques de chaque position de l'alignement multiple : une substitution de valine en tyrosine peut alors avoir un score différent à la position *i* et à la position *j* (**figure 13**).

Une PSSM est construite à partir d'un alignement multiple : la distribution des acides aminés et des insertions dans chaque colonne de l'alignement multiple permet d'extraire une fréquence d'occurrences pour chaque position qui peut-être traduite en un vecteur de scores de probabilités. Pour obtenir une PSSM plus précise, il est possible d'enrichir les fréquences observées dans l'alignement multiple par la connaissance que l'on a *a priori* des relations entre acides aminés en utilisant soit les matrices de substitutions classiques comme PAM et BLOSUM (Dayhoff, 1973; Dayhoff, 1978; Henikoff and Henikoff, 1992; Kosiol and Goldman, 2005), soit un modèle plus fin basé sur les mélanges de distributions de probabilités de Dirichlet (Sjolander et al., 1996).



P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	E	0	-2	2	1	1	2	0	-4	1	0	0	0	-3	0	0	-3	0	-1	-1	2
2	P	-1	0	0	0	1	0	-3	-4	4	-4	0	-1	0	-1	0	2	2	1	0	0
3	E	-1	0	1	0	1	0	0	-4	2	0	-1	0	1	-1	0	-3	1	0	0	2
4	G	0	3	-1	-1	-1	1	-4	-5	3	-4	0	-1	-4	2	0	-3	0	-3	2	0
5	S	0	-3	0	-1	0	1	-2	-5	0	-4	2	-2	-3	1	0	1	2	0	1	2
6	T	-2	0	1	0	0	0	1	-3	-2	-4	0	2	3	1	-2	1	1	-3	1	0
7	V	0	-4	1	0	2	2	2	1	1	1	1	0	0	-1	-2	0	0	0	0	0
8	D	1	0	-1	-4	-2	0	-5	-5	0	3	0	-1	-4	3	0	-3	2	0	3	2
9	V	1	0	1	0	5	0	-3	-4	-4	5	-3	0	-3	-1	-1	-4	-4	-4	-4	-1
10	V	0	-2	2	0	2	1	2	-2	-4	-3	0	0	4	-1	-1	0	-1	0	-3	0
P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	V	-1	-2	4	1	5	0	0	1	0	-3	-3	-2	0	-4	-4	-4	-4	-4	-5	-4
12	S	-1	-3	-2	0	1	2	0	-4	-1	1	4	4	-3	-2	0	-3	-1	0	0	-2
13	H	0	-3	0	0	0	0	2	1	-4	3	-1	-2	1	1	0	5	0	0	1	1
14	I	1	-2	5	0	4	1	2	1	-4	2	-2	-1	1	-1	-4	-4	-4	-4	-4	-4
15	D	0	-1	0	0	1	-2	0	-4	0	1	-1	0	0	2	0	-2	0	-2	3	2
16	S	-2	2	-1	-2	-1	-3	-4	-5	-3	0	4	0	-3	2	0	4	0	-2	2	0
17	P	0	0	-2	0	-1	-3	-1	-5	6	2	0	1	-4	0	0	0	-3	-3	2	-2
18	G	0	3	-3	0	0	1	0	-4	-1	3	2	0	0	2	0	-3	0	-3	-3	0
19	E	-3	-1	-1	-1	-2	-3	-3	-4	0	-4	0	1	1	1	0	4	1	2	2	2
20	F	-3	-2	3	0	2	-1	7	2	-5	0	-4	-1	0	-5	-5	-4	-5	-5	-5	-5
P	C	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E

figure 13 : La figure représente la matrice PSSM des 20 premières positions de famille des domaines Tudor (à l'aide du programme PSSM Viewer en utilisant le profil PFam 00567 pour les domaines Tudor). La première et la deuxième colonnes indiquent respectivement la position dans l'alignement multiple et l'acide aminé consensus à cette position. Par la suite, on alignera toutes les nouvelles séquences sur cette séquence consensus, en utilisant les vecteurs de scores de substitution indiqués à droite. Par exemple, la substitution V→Y sera défavorable à la position 9 (score = -3), et favorable à la position 10 (score = 4).

Les scores stockés dans une PSSM sont généralement des entiers positifs ou négatifs (voir **figure 13**). Une valeur positive indique que cette substitution est surreprésentée au sein de l'alignement multiple tandis qu'une valeur négative traduit le contraire. On peut en conclure que les positions où l'on trouve des valeurs positives élevées sont celles soumises à des pressions de sélection particulières associées par exemple au repliement, aux sites actifs ou aux surfaces d'interaction intra ou inter-moléculaires.

Les PSSM, ou profils, ont été introduit au sein du programme PSI-BLAST (Altschul et al., 1997) qui a révolutionné la détection d'homologie à faible identité de séquence. Par la suite, ce formalisme a été utilisé dans de nombreux programmes d'alignement multiple ou de recherche de nouvelles séquences homologues (Gowri et al., 2006; Kann et al., 2005; Marchler-Bauer et al., 2002; Schaffer et al., 1999).

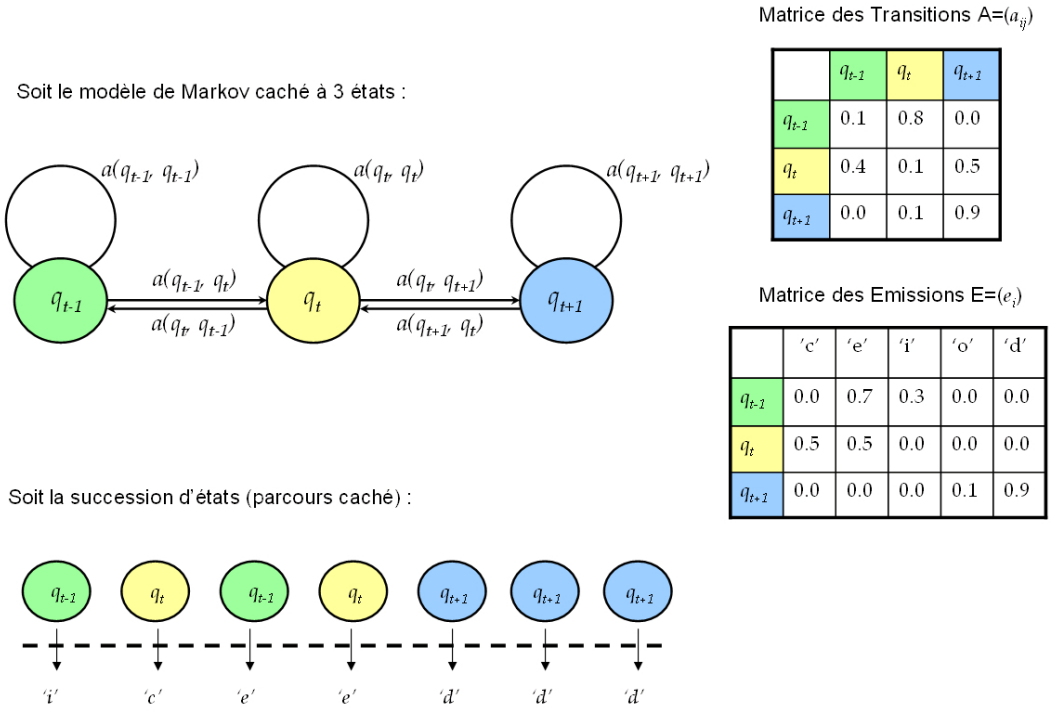
Les modèles de Markov cachés (HMM). Les modèles de Markov cachés, ou automates stochastiques à états cachés, constituent un formalisme statistique puissant dont le domaine d'application est très vaste (reconnaissance d'images, d'empreintes digitales, de langage, intelligence artificielle, bioinformatique, etc). Ils ont été introduits dans les années 1960 par Léonard Baum, et reposent sur une propriété fondamentale : l'état de l'automate à l'instant *t* dépend uniquement de son état à l'instant *(t-1)*.

Formellement (illustration **figure 14**), un modèle de Markov caché est défini comme un quadruplet (Q, Π, A, E) tel que :

- Q est l'ensemble des états $q_1,q_2….q_M$ de l'automate ;
- $\Pi = \pi_1...\pi_M$ est un vecteur définissant pour chaque état la probabilité qu'il soit l'état initial ;
- A = (a_{ij}) est une matrice MxM définissant les probabilités de transitions d'un état *i* vers un état *j* et telle que pour tout *i*, $\sum_{j=1}^{M} a_{ij} = 1$;
- E = (e_{ik}) est une matrice MxN définissant pour chaque état *i* la probabilité d'émettre le symbole *k* et telle que pour tout *i*, $\sum_{k=1}^{N} e_{ik} = 1$.

Intuitivement, on comprend que l'observation dont on dispose est la séquence émise par une suite d'états, mais cette suite d'états reste « cachée » et doit être déduite des observations à l'aide d'algorithmes adéquats. C'est la raison pour laquelle ces automates stochastiques sont dits « à états cachés ».

Dans le cadre de la bioinformatique et plus précisément de l'alignement des séquences, l'idée est de construire un modèle de Markov caché $\mathcal{H}$ représentant l'alignement multiple, c'est-à-dire tel que toutes les séquences constituant l'alignement multiple soient des observations issues d'un parcours de $\mathcal{H}$. En 1996, Sean Eddy (*Washington University School of Medecine*, Saint Louis, USA) a introduit des modèles de Markov cachés particulièrement adapté à la modélisation d'alignements multiples, et utilisés de nos jours au sein des programmes HMMER et SAMT2K (Eddy, 1996; Eddy, 1998; Karplus et al., 1998; Karplus et al., 2005).



	q_{t-1}	q_t	q_{t+1}
q_{t-1}	0.1	0.8	0.0
q_t	0.4	0.1	0.5
q_{t+1}	0.0	0.1	0.9

	'c'	'e'	'i'	'o'	'd'
q_{t-1}	0.0	0.7	0.3	0.0	0.0
q_t	0.5	0.5	0.0	0.0	0.0
q_{t+1}	0.0	0.0	0.0	0.1	0.9

figure 14 : Modèle de Markov caché contenant 3 états : q_{t-1}, q_t, q_{t+1}. La matrice de transition A=(a_{ij}) indique les probabilités de transition de chaque état *i* vers chaque état *j*. La somme des probabilités de transition sortantes de chaque état (somme des éléments d'une ligne) est égale à 1. La matrice des émissions stocke, pour chaque état, la probabilité d'émettre chacune des 5 lettres 'c','e','i','o','d', dont la somme est égale à 1. Pour une succession d'états (q_{t-1},q_t,q_{t-1},q_t,q_{t+1},q_{t+1},q_{t+1}), on peut observer une séquence ('i','c','e','e','d','d','d').

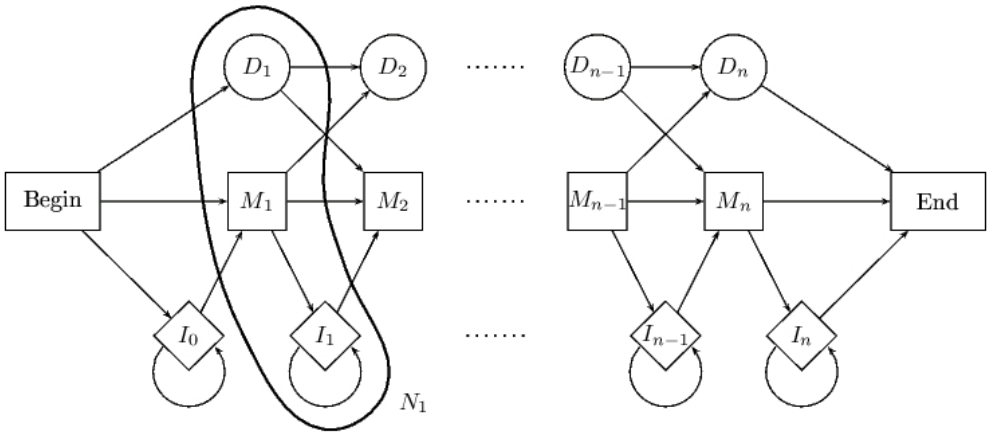


figure 15 : Illustration du Plan7 utilisé au sein de HMMer; chaque triplet N_i = (D_i, M_i, I_i) représente un nœud du HMM.

Chaque colonne p de l'alignement multiple est modélisée par un état dont la probabilité d'émettre un acide aminé donné est dépendante de la composition en acides aminés de la colonne p, que l'on peut pondérer avec des matrices de substitution ou un mélange de distributions de probabilités de Dirichlet (Brown et al., 1993; Dayhoff, 1973; Dayhoff, 1978; Henikoff and Henikoff, 1992; Kosiol and Goldman, 2005; Sjolander et al., 1996). Ces états qui émettent des acides aminés selon des probabilités spécifiques à chaque position sont les états d'appariements notés M_i. Pour modéliser les insertions et délétions, on utilise des états dédiés. Ainsi, un état de délétion D_i permet de passer directement de M_{i-1} à M_{i+1}, et donc de n'émettre aucun acide aminé en M_i. Les insertions s'intercalent quant à elles entre deux états d'appariement consécutifs M_i et M_{i+1} et permettent d'émettre un ou plusieurs acide(s) aminé(s). Selon le plan utilisé, toutes les transitions entre états ne sont pas autorisées. Notamment, il n'existe pas de transition d'une insertion vers une délétion et réciproquement dans le Plan7 (**figure 15**), alors que ces transitions sont autorisées au sein du Plan9.

Ajouter une nouvelle séquence s_{obs} au sein d'un alignement multiple *MSA* revient à rechercher la séquence d'états qui maximise la probabilité d'émission de s_{obs} au sein du modèle de Markov caché représentant *MSA*. L'intérêt d'utiliser ce formalisme des modèles de Markov cachés est que de nombreux algorithmes ont déjà été développés pour traiter des problèmes classiques. Parmi ceux-ci, l'algorithme mis au point par Andrew Viterbi en 1967, permet, connaissant tous les paramètres du modèle de Markov caché $\mathcal{H}$=(Q,π,A,E), et la séquence observée s_{obs}, de déterminer la séquence d'états cachés telle que la probabilité d'émission de s_{obs} soit maximale (Forney, 1973; Viterbi, 1967). De fait, cet algorithme est parfaitement adapté à la recherche de l'alignement optimal entre une séquence et un alignement multiple donné.

L'utilisation de l'algorithme de Viterbi permet donc de résoudre le problème des alignements séquence-HMM. Des travaux ont montré que les alignements produits par cette approche sont de meilleure qualité que les alignements séquence-profil construits en utilisant les PSSM (Eddy, 1998; Karplus and Hu, 2001; Krogh et al., 1994).

1.4.7 Les alignements profil-profil et HMM-HMM.

Dans le cadre de la détection d'homologies lointaines entre séquences, la comparaison de profil à profil, ou de HMM à HMM, a permis d'atteindre des résultats supérieurs à ceux obtenus avec les approches évoquées précédemment. Ces méthodes ont notamment été appliquées dans le but d'identifier des relations évolutives inattendues entre des familles de protéines (Kunin et al., 2001; Pietrokovski, 1996; Sadreyev and Grishin, 2003) ou lors du concours CASP5 pour trouver une structure de référence partageant moins de 6% d'identité de séquences avec la séquence à modéliser (Tramontano and Morea, 2003). De par leur efficacité à rechercher des homologies lointaines, ces méthodes permettent un élargissement important de l'espace des cibles potentielles de la modélisation comparative. Parmi les différents programmes développés ces dernières années, on distingue d'une part les programmes comparant deux profils (Ginalski and Rychlewski, 2003; Ginalski et al., 2004; Pietrokovski, 1996; Rychlewski et al., 2000; Sadreyev and Grishin, 2003; Yona and Levitt, 2002) et d'autre part deux programmes plus récents COACH et HHsearch, qui comparent respectivement un profil à un HMM et deux HMM entre eux (Edgar and Sjolander, 2004; Soding, 2005).

Dans le cadre des comparaisons profil à profil, des scores one été introduits comme par exemple ceux basés sur la différence d'entropie entre les deux distributions pour évaluer l'adéquation d'un alignement entre deux positions (Sadreyev and Grishin, 2003; Yona and Levitt, 2002). La comparaison mixte entre un profil et un HMM est quant à elle réalisée en se référant à la probabilité pour le HMM d'émettre la séquence consensus du profil (Edgar and Sjolander, 2004). Enfin, dans le cadre de l'alignement de deux HMMs, l'algorithme dédié cherche à maximiser les probabilités de co-émission des acides aminés (Soding, 2005).

Au cours des dernières années, quelques travaux dans les équipes d'Adam Godzik (*Burnham Institute*, La Jolla, USA) et David Baker (*Howard Hugues Medical Institute*, Seattle, USA) ont introduit les méthodes permettant de générer des alignements profil-profil alternatifs (Chivian and Baker, 2006; Jaroszewski et al., 2002). Ces méthodes sont toutes basées sur des approches paramétriques, parfois couplées au principe d'*Iterative Elimination Method* développé par Saqi et Sternberg dans le cas des alignements par paires. Jusqu'à présent et à

notre connaissance, aucune méthode permettant de générer le voisinage optimal des alignements profil-profil ou HMM-HMM n'a été développée.

1.4.8 Autres techniques de prédiction de structure intégrant de façon explicite l'information structurale.

Parmi les stratégies développées pour améliorer la qualité des alignements à basse identité de séquence, le *threading* (enfilage) occupe une place importante. Sa spécificité consiste à intégrer des informations structurales de façon explicite ; ainsi la séquence à modéliser n'est plus alignée sur une ou plusieurs autre(s) séquence(s) de structure(s) connue(s), mais directement sur la structure elle-même. L'idée est « d'enfiler » la séquence à modéliser sur la structure modèle pour identifier la correspondance séquence-structure la plus vraisemblable. Une fois cette correspondance séquence-structure optimisée, celle-ci est projetée en terme d'alignement des séquences pour continuer le processus classique de modélisation comparative.

Au cours des années 1990-2000, ces méthodes d'enfilage ont connu un essor très rapide (Jones, 1999; Kelley et al., 2000; McGuffin and Jones, 2003; Shi et al., 2001). Leur capacité à détecter des homologies lointaines en intégrant des informations structurales a été soulignée au cours des concours CASP (Moult, 2005).

Néanmoins, la complexité algorithmique du problème de l'alignement séquence-structure induit l'utilisation de certaines approximations qui perturbent la qualité de l'alignement. Au final, le *threading* se révèle surtout une méthode de choix pour identifier le repliement d'une séquence de structure inconnue. Les performances similaires des méthodes de *threading* et de comparaison profil-profil au dernier concours CAFASP5 (2006) suggèrent que l'information structurale qui était explicitement intégrée dans les algorithmes de *threading* se trouve désormais efficacement intégrée de façon implicite dans les méthodes de comparaison profil-profil. Etant donné la flexibilité accrue des méthodes de comparaison profil-profil (rapidité, structure 3D non requise), il est probable que dans les années à venir elles dominent le développement de nouvelles stratégies de type *threading*.

1.5 Optimisation du placement des chaînes latérales sur un squelette fixe.

1.5.1 Description du problème SCP.

Le problème, connu sous le nom de SCP pour *Side-Chain Positionning* est le suivant : ne connaissant que le positionnement du squelette peptidique et la séquence protéique correspondante, comment déduire le positionnement optimal des chaînes latérales ?

Il est communément admis que pour chaque acide aminé, il existe un nombre fini de conformations représentatives (**figure 16**) appelées rotamères par contraction de « *rotational isomers* ». En considérant que le repliement d'une protéine correspond à l'état dans lequel l'énergie libre est minimale, SCP consiste donc à rechercher la combinaison de rotamères qui minimise l'énergie libre.

figure 16 : Illustration de la notion de rotamère avec l'exemple de la phénylalanine. A gauche, une phénylalanine dans un rotamère donné. A droite, les 4 rotamères les plus abondants de la phénylalanine.

1.5.2 Définition des angles dièdres caractérisant le squelette peptidique et les chaînes latérales.

Dans l'espace tridimensionnel, on définit un angle dièdre χ (ou angle de torsion) entre 4 points *a*, *b*, *c* et *d* par la formule suivante :

$$\chi = sign\left[\vec{cb}\cdot\left(\vec{ab}\times\vec{cb}\right)\times\left(\vec{cb}\times\vec{cd}\right)\right]\arccos\frac{\left(\vec{ab}\times\vec{cb}\right)\cdot\left(\vec{cb}\times\vec{cd}\right)}{\left|\vec{ab}\times\vec{cb}\right|\left|\vec{cb}\times\vec{cd}\right|}$$

La conformation d'un acide aminé au sein d'une protéine peut être décrite par un ensemble d'angles dièdres : les angles ϕ, ψ et ω positionnent squelette peptidique et les angles χ_1, χ_2 … χ_5 positionnent la chaîne latérale (**figure 17**-A-B). Pour le squelette peptidique, l'angle ϕ

caractérise plus précisément la torsion autour de l'axe N-Cα et l'angle ψ la torsion autour de l'axe Cα-C. L'angle ω de la torsion autour de l'axe C-N tend à être planaire car le caractère partiellement double de la liaison peptidique empêche la rotation autour de la liaison C-N (**figure 17**-C). Les angles χ_1, χ_2 … χ_k caractérisent les angles de torsion le long de la chaîne latérale. Le nombre d'angles χ nécessaires pour décrire la conformation d'une chaîne latérale dépend du résidu étudié : il varie de cinq pour l'arginine à zéro pour l'alanine et la glycine.

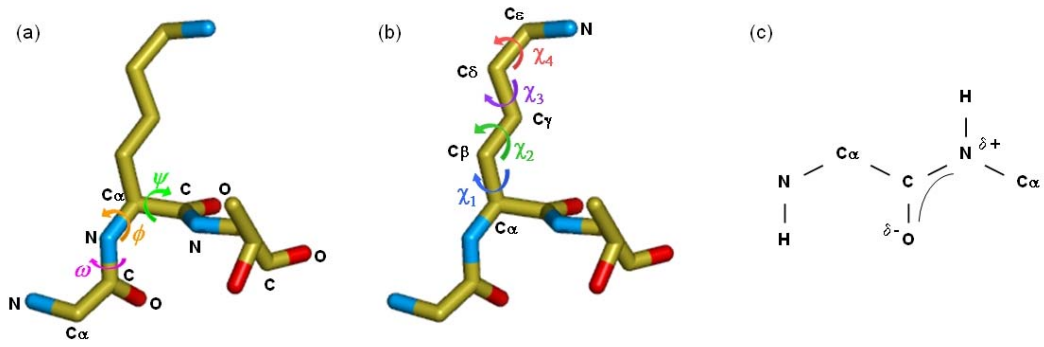


figure 17 : Illustration des angles dièdres définissant la conformation d'un résidu sur un court peptide GKT (glycine – lysine – thréonine). **(a)** Les angles ϕ et ψ caractérisent la position du squelette peptidique. **(b)** Les angles χ_1, χ_2, χ_3 et χ_4 caractérisent la position de la chaîne latérale (de la lysine dans l'exemple). **(c)** La liaison du carbone carbonyle C avec l'azote N dans la liaison peptidique est plus courte que la liaison simple C-N mais plus longue qu'une liaison double C=N classique. Le caractère partiellement double de la liaison peptidique empêche la rotation autour de la liaison C-N : le groupe peptidique est confiné dans un plan.

1.5.3 Approches heuristiques existantes.

Le premier programme développé pour prédire la conformation des chaînes latérales sur un squelette fixe a été présenté par Roland Dunbrack et Martin Karplus en 1993 (Dunbrack and Karplus, 1993). Depuis, de nombreux autres programmes ont été développés. Ces programmes diffèrent les uns des autres principalement par trois facteurs : (i) la bibliothèque de rotamères, (ii) l'algorithme d'optimisation, (iii) la fonction d'énergie.

Facteur lié à la bibliothèque de rotamères. De nombreuses bibliothèques de rotamères ont été introduites au cours des 20 dernières années. On distingue trois catégories de bibliothèques de rotamères : les bibliothèques de rotamères dépendantes du squelette peptidique (BBDEP), les bibliothèques de rotamères dépendantes des structures secondaires (SSDEP), et enfin les bibliothèques de rotamères indépendantes du squelette peptidique

(BBIND). Dans les bibliothèques de rotamères dépendantes du squelette peptidique, les rotamères testés pour un résidu sont fonction des angles ϕ et ψ du résidu. Dans les bibliothèques de rotamères dépendantes des structures secondaires, les rotamères testés ne sont pas les mêmes selon que l'on se trouve dans une hélice α, un feuillet β ou dans une région non structurée. A l'opposé, avec les bibliothèques de rotamères indépendantes du squelette peptidique, les rotamères testés ne dépendent pas de la conformation locale du squelette. La granulosité des différentes bibliothèques et le nombre total de rotamères varient également beaucoup d'une bibliothèque à l'autre.

Facteur Algorithmique. SCP est un problème combinatoire particulièrement difficile. En effet, Niles Pierce et Erik Winfree ont démontré en 2002 que SCP appartient à la classe des problèmes NP-complets (Pierce and Winfree, 2002).

La classe des problèmes NP-complets rassemble les problèmes tels que :

- Le problème appartient à la classe NP, ce qui signifie qu'on ne sait pas le résoudre au moyen d'un algorithme efficace (cad. polynomial en temps). Plus intuitivement, les problèmes dans NP sont tous les problèmes qui ne peuvent être résolus qu'en énumérant l'ensemble des solutions possibles.
- S'il existait un algorithme efficace pour l'un d'entre eux, alors on pourrait en déduire un algorithme efficace pour l'ensemble des problèmes de la classe. Néanmoins les mathématiciens s'accordent à penser qu'un tel algorithme n'existe pas.

Devant la complexité du problème SCP global, différentes stratégies ont été adoptées. Certains groupes ont choisi de ne pas chercher le minimum d'énergie libre global et de se contenter d'un minimum local (Mendes et al., 2001; Peterson et al., 2004; Tuffery et al., 1997; Xiang and Honig, 2001). D'autres groupes ont préféré ne pas optimiser l'ensemble des résidus de la protéine en une seule étape mais plutôt optimiser indépendamment des sous-ensembles de résidus (Canutescu et al., 2003).

Facteur lié à la fonction d'énergie. Pour chaque combinaison de rotamères générée, il faut tester si celle-ci est énergétiquement plus favorable que les combinaisons précédentes. Il s'agit d'approximer l'énergie libre associée à cette combinaison de façon précise mais très rapide car le nombre total de combinaisons à tester est très important. Les équipes ont choisi des stratégies différentes en privilégiant soit la rapidité, comme c'est la cas dans certains

programmes pour lesquels la fonction d'énergie est réduite à une estimation du terme de Van der Waals (Canutescu et al., 2003; Xiang and Honig, 2001), soit la précision en utilisant des fonctions d'énergie issues de champs de force et souvent couplées à des termes statistiques (Mendes et al., 2001; Peterson et al., 2004).

1.6 Fonctions de score développées pour le design automatique et semi-automatique de structures.

1.6.1 Introduction au problème du *design*.

Le *design* de protéines assisté par ordinateur est un procédé visant à modifier la séquence d'une protéine afin d'améliorer certaines de ses propriétés telles que sa stabilité, sa fonction, ou sa spécificité d'interaction. La prédiction des mutations à effectuer découle d'une analyse bioinformatique de la structure de la protéine. En se basant sur les méthodes de *design*, les différents succès reportés dans la littérature illustrent les champs d'applications de ces approches. En effet, plusieurs études ont montré que la stabilisation de protéines basée sur leurs structures pouvait être envisagée efficacement en utilisant des approches automatiques de *design* (Dahiyat and Mayo, 1997; Filikov et al., 2002; Korkegian et al., 2005; Ventura et al., 2002). Une autre application de ces approches, l'ingénierie de nouveaux repliements protéiques, représente un réel défi des méthodes de *design*. À l'heure actuelle, un seul programme a réussi à relever ce défi, le programme ROSETTA. À partir de ce programme, une nouvelle protéine de 93 résidus, baptisée Top7, présentant une séquence et une topologie originales, a récemment été synthétisée (Kuhlman et al., 2003).

Les progrès réalisés dans ce domaine montrent que le succès des méthodes de *design* repose sur deux facteurs : des algorithmes efficaces permettant de gérer l'exploration de l'espace des séquences et sa combinatoire exponentielle, et des fonctions d'énergie (ou d'évaluation) dont le rôle est de trier les séquences relativement à leur adéquation avec le repliement de la protéine. Du fait de la précision requise par ce genre d'approche leur efficacité dépend directement des fonctions d'énergie sur lesquelles elles s'appuient.

1.6.2 Trois catégories de fonctions d'énergie.

Le développement de fonctions d'évaluation précises est l'un des enjeux majeurs de la bioinformatique structurale, particulièrement en ce qui concerne le *design* de protéines. A l'heure actuelle, différents programmes de *design* des protéines ont été rendus accessibles à la communauté scientifique (**table 4**). Les fonctions d'énergies utilisées par ses programmes se divisent en trois sous catégories (Lazaridis and Karplus, 2000) : les méthodes statistiques (SEEF pour *statistical effective energy function*), les méthodes basées sur un champ de force physique (PEEF pour *physical effective energy function*) et enfin une troisième classe de méthode basée sur l'utilisation de données expérimentales (EEEF pour *empirical effective energy function*).

Méthodes SEEF. Les potentiels statistiques sont dérivés des structures de la PDB. Leur principe est d'analyser la fréquence de certaines interactions dans la PDB et de les convertir en énergies. On peut citer comme exemple de potentiel statistique, PopMusic (Rooman …) qui a été utilisé pour stabiliser certaines protéines. Il s'agit donc d'un potentiel statistique prenant en compte les distances entre résidus et également leur conformation locale. La caractéristique commune à toutes ces méthodes SEEF est de considérer une représentation simplifiée des protéines. De ce fait, les méthodes SEEFs ont l'avantage d'être très rapides et peu sensibles aux petites erreurs de positionnement des atomes ; c'est la raison pour laquelle elles sont fréquemment utilisées pour l'évaluation des modèles obtenus par homologie (Lazaridis and Karplus, 2000).

Méthodes PEEF. Les potentiels physiques combinent des fonctions d'énergies issues de la mécanique moléculaire, ainsi que des modèles prenant en compte les effets de la solvatation sur l'énergie libre du système. La principale différence par rapport aux méthodes statistiques concerne la paramétrisation de ces méthodes qui n'est pas dérivée de la structure de protéines, mais de la mesure de paramètres physiques. Ces méthodes sont largement utilisées dans le domaine de la dynamique moléculaire, pour simuler le comportement des protéines. Cependant, ces méthodes semblent moins adaptées aux problèmes de prédiction de structures des protéines, compte tenu du temps de calcul qu'elles nécessitent.

Type	Méthode	Description	Site web	Référence
SEEF	PoPMuSiC	Prédiction de l'effet stabilisant de mutations ponctuelles Potentiel statistique issu de l'analyse de bases de données	http://babylone.ulb.ac.be/popmusic	[1]
SEEF	I-Mutant	Prédiction de l'effet stabilisant de mutations ponctuelles Version 1.0 basée sur un réseau de neurones Version 2.0 basée sur un SVM Apprentissage à partir de la base de données Protherm[ref]	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi	[2]
PEEF	EGAD	Ingénierie des protéines Prédiction de l'effet stabilisant de mutations ponctuelles Basée sur le champ de force OPLS-AA	http://egad.berkeley.edu/software.php	[3]
EEEF	Dfire-Dmutant	Prédiction de l'effet stabilisant de mutations ponctuelles Potentiel distance dépendant, résidu-spécifique	http://sparks.informatics.iupui.edu/	[4]
EEEF	Foldx	Prédiction de l'effet stabilisant de mutations ponctuelles Evaluation de l'énergie libre ΔG Paramétrisation à partir de mutants expérimentaux	http://foldx.embl.de	[5]
EEEF	Rosetta	Ingénierie des protéines Prédiction de l'effet stabilisant de mutations ponctuelles Evaluation de l'énergie de la protéine cible	http://rosettadesign.med.unc.edu	[6]

table 4 : Table récapitulative des différentes méthodes de design. Les références sont [1] Gilis et al. 2000 ; [2] Capriotti et al. 2004, Capriotti et al. 2005 ; [3] Pokala et al. 2005 ; [4] Zhou et al. 2002 ; [5] Guerois et al. 2002 ; [6] Kuhlman et al. 2003.

Méthodes EEEF. Ces méthodes combinent une description physique des interactions et une connaissance basées sur des données expérimentales et statistiques. À titre d'exemples, l'algorithme AGADIR (Munoz and Serrano, 1995), le programme FOLD-X (Guerois et al., 2002), ainsi que le programme ROSETTA sont trois méthodes empiriques couramment utilisées pour le *design* de mutation(s) permettant d'augmenter la stabilité de protéines. En particulier, différents succès de *design* protéique ont été reportés aussi bien par le programme FOLD-X (van der Sloot et al., 2004) que par le programme ROSETTADESIGN (Dantas et al., 2003; Kuhlman et al., 2003; Kuhlman et al., 2001; Kuhlman et al., 2002; Nauli et al., 2001). Ces programmes sont présentés plus en détails dans la section suivante.

1.6.3 Fonctions d'énergie empiriques pour le design : Foldx et RosettaDesign

FOLDX. FOLDX est développé dans le groupe de Luis Serrano à l'EMBL, Heidelberg. La fonction d'énergie de FOLDX, FOLDEF, est un champ de force empirique destiné à évaluer les effets d'une mutation sur la stabilité de protéines ou d'acides nucléiques, en se basant sur leur structure. La fonction d'énergie FOLDEF évalue l'énergie libre ΔG d'une protéine par la combinaison de différents termes :

$$\Delta G = W_{vdw}\Delta G_{vdw} + W_{solvH}\Delta G_{solvH} + W_{solvP}\Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + W_{mc}T\Delta S_{mc} + W_{sc}T\Delta S_{sc}$$

a) ΔG_{vdw} est la somme des contributions de Van der Waals de tous les atomes du système
b) ΔG_{solvH} et ΔG_{solvP} représentent la différence d'énergie de solvatation entre l'état déplié et l'état replié, pour les groupements apolaires et polaires respectivement
c) ΔG_{hbond} représente la différence d'énergie libre entre la formation d'une liaison hydrogène intra-moléculaire et la formation d'une liaison hydrogène avec le solvant
d) ΔG_{wb} pour la prise en compte des molécules d'eau enfouies
e) ΔG_{el} représente les contributions électrostatiques entre groupements chargés
f) ΔS_{mc} représente le coût entropique relatif au repliement du squelette peptidique
g) ΔS_{sc} représente le coût entropique relatif aux conformations des chaînes latérales

Les poids W_{vdw}, W_{solvH}, W_{solvP}, W_{mc} and W_{sc} ont été optimisées à partir de données expérimentales concernant 339 mutants ponctuels (Guerois et al., 2002). La capacité de

prédiction de la fonction FOLDEF a été récemment évaluée sur une base de données test de 1088 mutants ponctuels caractérisés expérimentalement. Le *design* de protéines par le programme FOLDX peut être réalisé en couplant cette approche à d'autres programmes de prédiction de structures, tel que le programme SCAP (Xiang and Honig, 2001).

ROSETTA. ROSETTA est une suite de programmes, incluant différents modules parmi lesquels ROSETTA AB INITIO, ROSETTANMR, ROSETTADESIGN, et ROSETTADOCK. Ces modules sont destinés respectivement à la prédiction de structure, à l'intégration de données RMN pour la résolution de structures, à l'identification de séquences optimales pour une structure protéique donnée, et à la prédiction de structure de complexes protéiques. La suite ROSETTA est développée par le laboratoire de David Baker de l'université de Washington (Seattle) et par un consortium organisé autour des anciens membres de ce laboratoire.

L'une des applications de cette suite, ROSETTADESIGN, permet l'ingénierie de protéines. L'objectif principal de ce programme est l'identification de séquences protéiques permettant la stabilisation de structures protéiques cibles. À partir de la structure d'une protéine cible, et de la position des résidus à optimiser, ROSETTADESIGN détermine spécifiquement les mutations permettant d'augmenter la stabilité de la protéine. Pour cela, ROSETTADESIGN couple une fonction d'énergie d'évaluation, et une procédure d'optimisation destinée à l'exploration de l'espace des séquences.

La fonction d'énergie de ROSETTADESIGN est constituée de différents termes :

$$E_{protein} = W_{rot}E_{rot} + W_{atr}E_{atr} + W_{rep}E_{rep} + W_{solv}E_{solv} + W_{pair}E_{pair} + W_{hbond}E_{hbond} - E_{ref}$$

(a) E_{atr} et E_{rep} correspondent respectivement à la composante attractive et répulsive du potentiel de Lennard-Jones.

(b) E_{solv} est l'énergie de solvatation calculée selon un modèle de solvant implicite, le modèle de Lazaridis-Karplus (Lazaridis and Karplus, 2000).

(c) E_{hbond} est le potentiel de liaison hydrogène basé sur les caractéristiques géométriques des liaisons hydrogènes de protéines issues de la PDB (Kortemme et al., 2003).

(d) E_{rot} est l'énergie associée à la conformation d'un rotamère donné. Cette énergie est basée selon des critères probabilistes issus de l'analyse des structures issues de la PDB.

(e) E_{pair} permet de modéliser les interactions électrostatiques entre résidus chargés. Ce terme est basé sur des critères statistiques issus d'analyses de la PDB (Simons et al., 1999).

(f) E_{ref} correspond à une énergie de référence, propre à chaque acide aminé.

Les poids W_{rot}, W_{atr}, W_{rep}, W_{solv} W_{pair} and W_{hbond} ainsi que les énergies de référence ont été optimisées à partir d'un jeu données expérimentales de 30 protéines (Kuhlman and Baker, 2000).

Concernant la phase d'optimisation, ROSETTADESIGN utilise un algorithme de recuit simulé. En partant d'une séquence protéique aléatoire, des mutations ponctuelles des résidus à modifier, couplées à des modifications de leur rotamères associés sont acceptées selon un critère de Métropolis. Les chaînes latérales des acides aminés à modifier vont ainsi adopter un ensemble discret de rotamères, tirés de la librairie de rotamères de Dunbrack (Dunbrack and Cohen, 1997). Certaines variations de ces conformations préférentielles sont autorisées pour les résidus enfouis au sein de la protéine, ceci par de légères modifications des valeurs de χ_1 et χ_2.

L'originalité du programme ROSETTADESIGN consiste en l'incorporation de la flexibilité du squelette peptidique dans le processus d'optimisation de séquence décrit précédemment. L'introduction de la flexibilité permet d'explorer un champ plus large de possibilités au niveau de l'espace des séquences, et représente donc une étape essentielle pour le design de protéines non naturelles de topologie inconnue.

Chapitre 2 : Détection et Modélisation des PRMs

Ce chapitre s'intéresse au problème de la détection et de la modélisation des « Peptide Recognition Modules ». La protéine humaine Nbs1 et son orthologue Xrs2 chez Saccharomyces cerevisiae constituent à ce titre deux exemples remarquables. Vincent Meyer, précédent doctorant au sein de notre laboratoire, a mis en évidence que ces deux protéines possèdent un tandem de domaines BRCT, inconnu jusqu'alors, sur une portion de la séquence fortement divergente. Toutefois, du fait de cette très forte divergence, la délimitation des domaines est imprécise. C'est la raison pour laquelle une étude approfondie sur la base d'une modélisation structurale a été entreprise.

2.1 Détection et Modélisation d'un tandem BRCT dans les protéines Nbs1 et Xrs2.

2.1.1 La protéine humaine Nbs1 et son orthologue Xrs2 chez la levure.

Chez l'homme, la protéine Nbs1 est un composant essentiel du complexe MRN comprenant les protéines Mre11, Rad50 et Nbs1 (Petrini and Stracker, 2003; van den Bosch et al., 2003; Zhang et al., 2006). Ce complexe protéique joue un rôle majeur dans la voie de signalisation des dommages de l'ADN en participant à la détection des cassures double-brins et en permettant le recrutement du complexe *Ataxia Telangiectasia Mutated* (ATM) à l'endroit précis de la lésion (Kobayashi et al., 2004; Lee and Paull, 2005). Il a été montré *in vitro* que des dimères ATM inactifs étaient activés en présence du complexe MRN, amenant ainsi une phosphorylation des kinases effectrices p53 et Chk2 (Lee and Paull, 2005).

La protéine Nbs1 comprend 754 résidus. Au sein de la séquence de Nbs1, différents domaines ont été repérés à partir de méthodes d'analyse de séquences et *via* des expériences biochimiques. La **figure 18** présente la composition de Nbs1 avec les délimitations des différents domaines. La région N-terminale comprend un domaine FHA immédiatement suivi d'un domaine BRCT. La partie C-terminale contient le site d'interaction reconnu par Mre11 (Desai-Mehta et al., 2001) et un motif de recrutement du complexe ATM (Falck et al., 2005).

La protéine Nbs1 tient son nom du syndrome de Nijmengen (NBS = *Nijmengen Breakage Syndrome*). Le syndrome NBS a longtemps été considéré comme une variante de l'ataxie (trouble de coordination) cérébelleuse. Sa prévalence est très hétérogène avec une fréquence plus élevée dans les pays de l'Est. Les patients atteints de NBS présentent généralement une microcéphalie associée à un retard mental, à un déficit immunitaire et à une hypofertilité. Ils ont de plus une prédisposition élevée à développer des cancers. Une délétion de 5 paires de bases provocant le clivage de la protéine Nbs1 en deux protéines plus petites, p26 et p70, est présente chez 95% des patients atteints de NBS (voir **figure 18**). La protéine p26, de 26kDa, inclut les positions 1-218 de Nbs1 et comprend les domaines FHA et BRCT de la région N-terminale. Grâce à la présence d'un site alternatif d'initiation de la traduction immédiatement en amont de la délétion de 5 paires de bases, la protéine p70, de 70kDa, correspondant à la partie C-terminale de Nbs1, est également synthétisée. Après une extension N-terminale de

18 résidus, la séquence de p70 est identique à celle de Nbs1 de la position 221 jusqu'à l'extrémité C-terminale de la protéine (Williams et al., 2002).

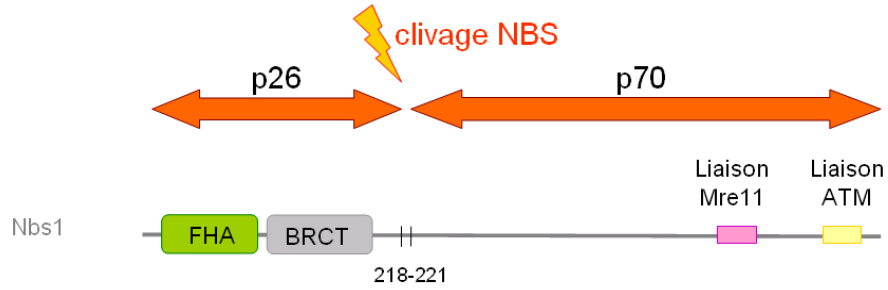


figure 18 : Organisation en domaines de la protéine humaine Nbs1. La région N-terminale comprend un domaine FHA (en vert) immédiatement suivi d'un domaine BRCT (en gris). La région C-terminale comporte les deux domaines de liaison à Mre11 (en rose) et au complexe ATM (en jaune). Chez les patients atteints du syndrome de Nijmengen, la protéine Nbs1 est coupée en deux au niveau d'une délétion de 5 paires de bases entre les positions 218 et 221. Il en résulte deux protéines : P26 (identique à la portion 1-218 de Nbs1), et P70 (après une insertion de 18 résidus, identique à la portion 221-754 de Nbs1).

L'homologue fonctionnel du complexe MRN chez *Saccharomyces cerevisiae* est le complexe MRX, incluant les protéines de levure Mre11, Rad50 et Xrs2, l'orthologue de Nbs1. Les techniques actuelles d'analyse de séquences ont permis d'identifier au sein de la séquence de Xrs2 un domaine FHA en N-terminal, comme au sein de Nbs1. L'existence de sites de liaison de Tel1, l'homologue d'ATM, et de Mre11 ont également été mis en évidence (**figure 19**). La composition en domaines des protéines Xrs2 et de Nbs1 semble donc légèrement différente puisque le domaine BRCT détecté dans Nbs1 n'a pas d'équivalent au sein de Xrs2.

figure 19: Organisation en domaines de la protéine de levure Xrs2. La région N-terminale comprend un domaine FHA (en vert). Les sites de liaison à Mre11 (en rose) et Tel1 (en jaune) sont situés dans la région C-terminale.

Dans la suite de cette section, nous montrons qu'en utilisant des techniques récentes de comparaison HMM-HMM, il est possible de détecter la présence d'un second domaine BRCT au sein de Nbs1. Malgré la forte divergence de séquence, un tandem BRCT a également pu être mis en évidence dans Xrs2. Ces identifications nous ont conduit à élaborer un modèle structural des domaines FHA et BRCT de Nbs1 et de Xrs2 (Becker et al., 2006) afin de valider l'analyse et explorer les implications biologiques de ces découvertes.

2.1.2 Détection d'un domaine BRCT caché.

Sur les 250 acides aminés suivant les domaines FHA de l'extrémité N-terminale, les séquences protéiques de Nbs1 et Xrs2 sont fortement divergentes (environ 10% d'identité). L'analyse manuelle que Vincent Meyer a effectuée au cours de sa thèse sur les protéines Nbs1 et Xrs2 a permis de détecter la présence d'un tandem BRCT immédiatement en aval du domaine FHA au sein de Nbs1 et Xrs2 (Meyer, 2007). La méthode employée pour détecter ces domaines est brièvement expliquée ci-dessous (pour plus de détails, voir l'**Article 1**).

Vincent Meyer a construit un profil initial regroupant les homologues proches de Nbs1 en effectuant une recherche PSI-BLAST (Altschul et al., 1997) sur une banque de séquences non redondante. Pour Xrs2, le profil initial a été construit en effectuant une recherche ciblée sur les protéines de levure. Ces profils ont ensuite été enrichis afin d'y inclure des séquences plus divergentes en s'appuyant sur une méthode d'alignements HMM-HMM nommée HHALIGN (Soding, 2005). Cette opération d'enrichissement a été répétée de façon itérative jusqu'à l'obtention d'un alignement multiple de 25 séquences allant de la protéine de levure Xrs2 jusqu'à la protéine humaine Nbs1. Avec cet alignement multiple de 25 séquences, un profil final a été construit et comparé à une banque de profils à l'aide du serveur HHPRED (Soding et al., 2005). Grâce à cette analyse, Vincent Meyer a pu montrer que les deux protéines Nbs1 et Xrs2 possédaient deux domaines BRCT en aval du domaine FHA N-terminal sur la portion de séquence très peu conservée.

La **figure 20** présente la composition en domaines des protéines Nbs1 et Xrs2 en prenant en compte l'existence des domaines BRCT « cachés ». Sur la partie N-terminale de Nbs1 et Xrs2, on remarque que les domaines FHA suivis des deux domaines BRCT s'enchaînent rapidement : les *linkers* (connecteurs) séparant les domaines deux à deux sont en effet très courts. Le domaine FHA n'est séparé du premier domaine BRCT que par 2 résidus ; cette longueur est constante dans toutes les espèces présentes au sein de notre alignement. Le *linker* séparant les deux domaines BRCT compte quant à lui environ 20 résidus. Curieusement, le clivage typique de Nbs1 chez les patients atteints de NBS se situe au niveau de ce *linker*.

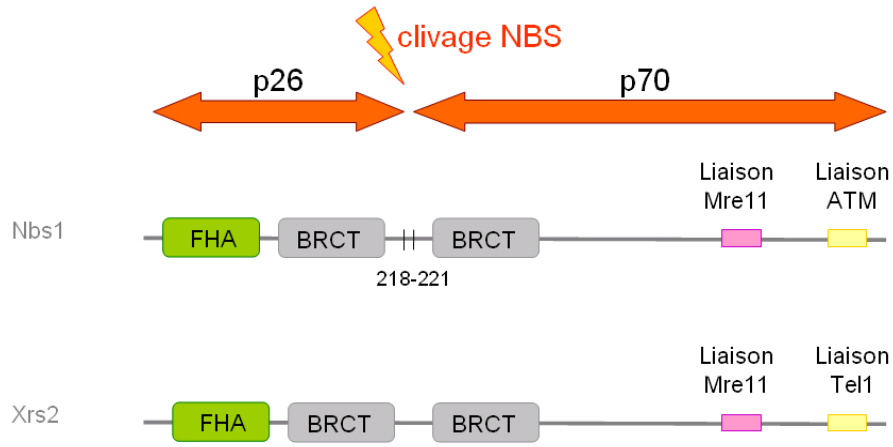


figure 20 : Composition en domaines des protéines Nbs1 et Xrs2 après l'identification des domaines BRCT « cachés ». Chacune des deux protéines comprend dans la région N-terminale un domaine FHA (en vert) suivi d'un tandem de domaines BRCT (en gris). La région C-terminale inclut les domaines de liaison à Mre11 et ATM pour Nbs1 et à et Mre11 et Tel1 pour Xrs2 (respectivement en rose et jaune). Le clivage typique de Nbs1 chez les patients atteints de NBS intervient au sein du court *linker* connectant les deux domaines BRCT.

2.1.3 Modélisation de la structure du tandem de domaines BRCT de Nbs1.

La partie du travail effectuée par Vincent Meyer a permis de détecter sans ambiguïté la présence de domaines BRCT très divergents au sein de Nbs1 et Xrs2. L'objectif suivant pour notre équipe était de produire expérimentalement ces domaines pour étudier leurs propriétés fonctionnelles. Toutefois, du fait de la très forte divergence de certaines régions de l'alignement, il subsistait un doute sur la délimitation exacte de l'extrémité C-terminale du tandem BRCT. Comme nous l'avons dit précédemment, la zone de 250 résidus qui suit immédiatement le domaine FHA et qui contient le tandem BRCT est très peu conservée (10% d'identité de séquence). Afin d'identifier le plus précisément possible les délimitations de ces domaines, j'ai effectué une étude approfondie s'appuyant sur la modélisation structurale du tandem. Cette étude a impliqué une exploration manuelle de nombreux alignements alternatifs et une évaluation des modèles structuraux correspondants par des potentiels statistiques et par des simulations de dynamique moléculaire. Cette approche itérative assez fastidieuse nous a néanmoins permis d'éliminer de nombreux alignements *a priori* possibles sur la base des alignements de séquences mais incompatibles dans le contexte de la structure tridimensionnelle.

En nous servant de l'alignement de séquence proposé directement par la comparaison HMM-HMM et ayant permis d'identifier le tandem de domaines BRCT, les premiers modèles du tandem domaines BRCT de Nbs1 ont été produits avec MODELLER8V2 (Sali and Blundell, 1993). L'évaluation de ces modèles par les potentiels statistiques VERIFY3D (Luthy et al., 1992), PROSA2003 (Sippl, 1993), PROQ et MAXSUB (Wallner and Elofsson, 2003) a révélé que plusieurs régions étaient probablement mal alignées et conduisaient à des structures peu probables. Pour VERIFY3D et PROSA2003, ses scores respectivement supérieurs à 0,1 et négatif permettent généralement de valider la fiabilité d'un modèle. Un score PROQ>1,5 couplé à un score MAXSUB>0,1 est un gage supplémentaire de fiabilité de la structure évaluée (Wallner and Elofsson, 2003). Les problèmes se concentraient dans la région C-terminale et plus particulièrement au niveau de la dernière hélice α (données non exposées). L'alignement initial a donc été raffiné par une exploration manuelle des différentes combinaisons possibles. Cette exploration centrée principalement sur la région C-terminale a conduit à un ensemble de dix alignements acceptables au vu des scores d'évaluation par les potentiels statistiques et relativement similaires les uns des autres.

Pour discriminer entre ces solutions, différentes simulations de dynamique moléculaire en solvant explicite ont été calculées en utilisant le programme GROMACS 3.2 et le champ de force OPLS (Van Der Spoel et al., 2005). Après avoir placé les modèles dans une boîte d'eau et introduit des ions pour neutraliser le système, une minimisation a été effectuée. La température du système a ensuite été progressivement amenée à 300K par un recuit simulé avant qu'une simulation de dynamique moléculaire de 5 nanosecondes ne débute. Pour la plupart des modèles, un dépliement progressif de la région C-terminale a été observé après les deux premières nanosecondes de simulation. L'origine du dépliement était de deux ordres : (i) longueurs des segments connectant les structures secondaires trop courtes et induisant des contraintes trop fortes sur la topologie du domaine (ii) existence de cavités qui ne pouvaient être compensées par la relaxation de la structure au cours de la simulation.

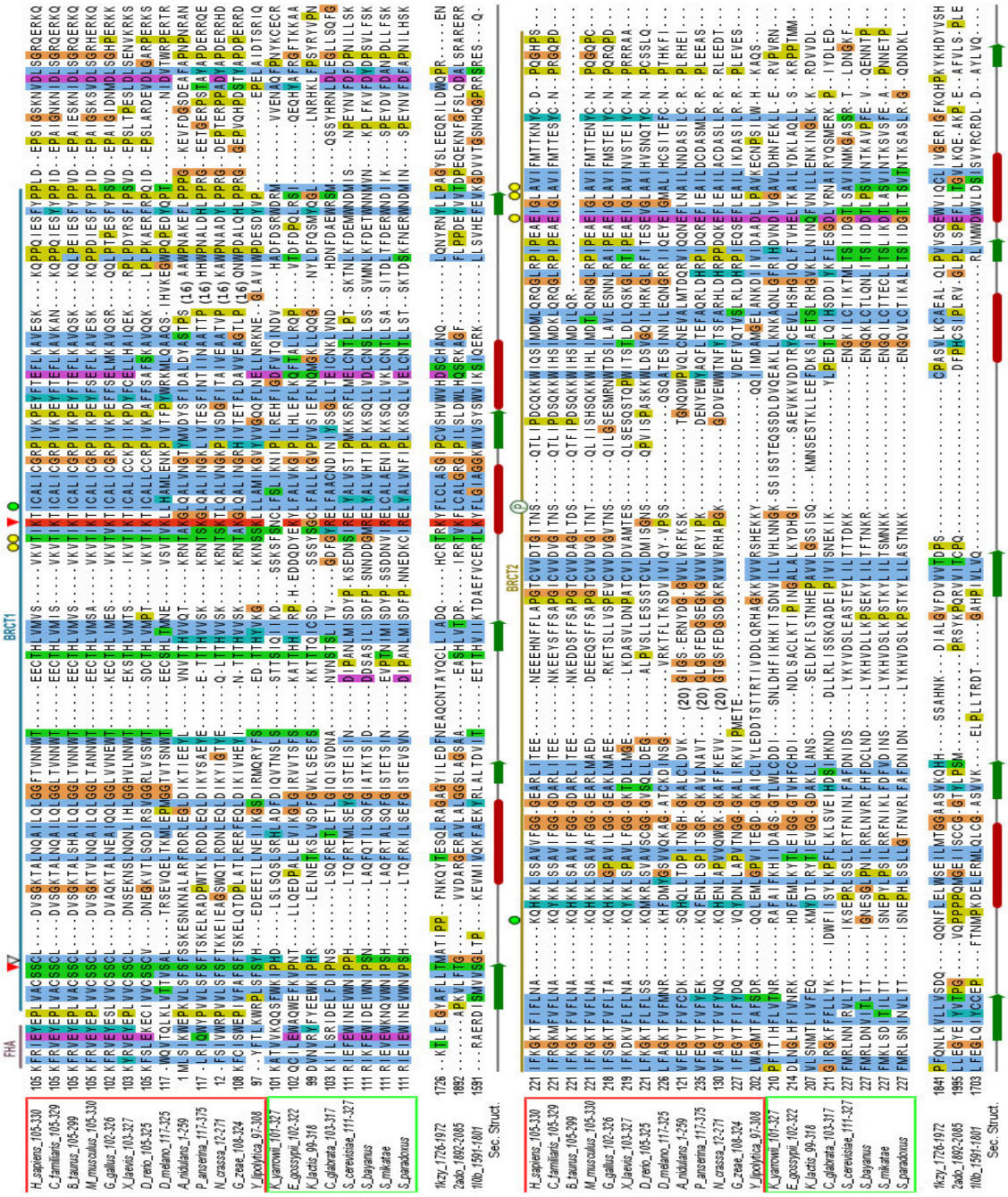


figure 21 : Alignement multiple des séquences du tandem BRCT de Nbs1, Xrs2, et d'autres séquences de tandems BRCT parmi lesquelles on trouve trois structures connues (1Kzy, 2Ado, 1L0b). Sur la ligne du bas est représentée la position des structures secondaires : les hélices en rouge et les brins en vert. La ligne du haut indique les délimitations du domaine BRCT1 et du domaine BRCT2. Les positions en contact avec le résidu pS dans les structures connues sont indiquées par des triangles rouges (le contact se fait via la chaîne latérale), et des triangles blancs (le contact se fait via le squelette). Les boîtes rouges et vertes encadrant les noms de séquences regroupent celles-ci en fonction du motif sur ces positions. Enfin, les positions en contact avec le résidu pS+3 sont indiquées par des cercles verts et celles en contact avec le reste du peptide par des cercles jaunes.

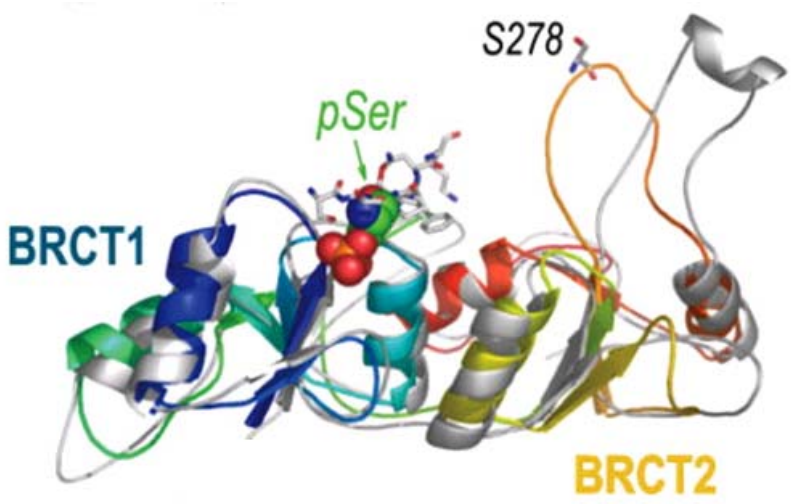


figure 22 : Modèle de la structure du tandem de domaines BRCT de Nbs1. Ce modèle a été produit par modélisation comparative *via* Modeller8v2 (Sali and Blundell, 1993) d'après les structures de références 2Ado et 1T15. La structure du tandem BRCT avant dynamique (en ruban, couleur arc-en-ciel), est superposée à celle après la simulation de dynamique moléculaire de 5ns (en ruban, gris). Le site connu pour reconnaître des phospho-sérines dans les autres tandems BRCT est indiqué (phospho-sérine en sphères, le peptide en bâtons).

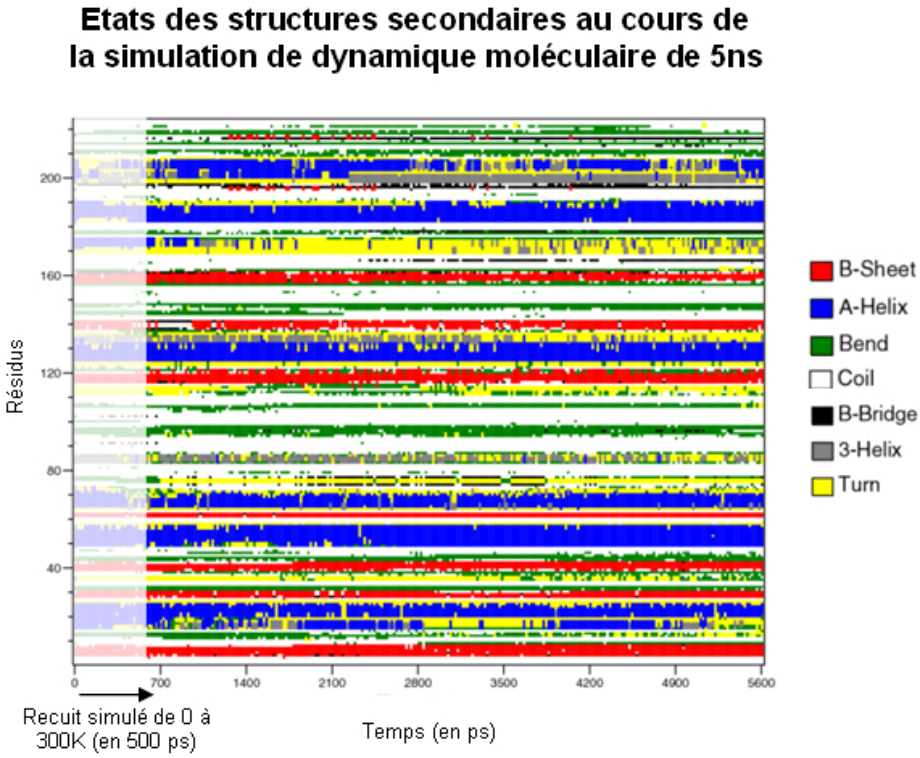


figure 23 : Stabilité des structures secondaires, calculées par DSSP (Kabsch et al., 1983), au cours des 5ns de simulation de dynamique moléculaire en solvant explicite. Les positions sont en ordonnée et le temps, en ps, en abscisse. Les premières 500ps (bande blanche) sont consacrées au recuit simulé ayant permis de chauffer le système de 0K à 300K. Les hélices α sont représentées par des points bleus, les brins β par des points rouges.

Pour un des alignements sélectionnés représenté dans la **figure 21**, nous avons observé que les structures secondaires de Nbs1 étaient toutes stables au cours de la dynamique (**figure 23**). Le modèle initial du tandem de domaines BRCT de Nbs1 ainsi que la structure obtenue après 5 ns de simulation sont superposés sur la **figure 22**. En s'appuyant sur cette solution d'alignement, un modèle du domaine de Xrs2 a également été généré et validé par les potentiels statistiques et par dynamique moléculaire. Les scores statistiques pour le modèle de Nbs1 sont (PROSA2003: –1,92), (VERIFY3D: 0,395), (PROQ : 3,51) et (MAXSUB : 0,348) et ceux du modèle de Xrs2 sont (PROSA2003 : –1,16), (VERIFY3D : 0,332), (PROQ : 3,75) and (MAXSUB : 0,338). La **figure 24** illustre les scores obtenus avec le potentiel VERIFY3D pour les deux modèles Xrs2 et Nbs1 le long de la séquence.

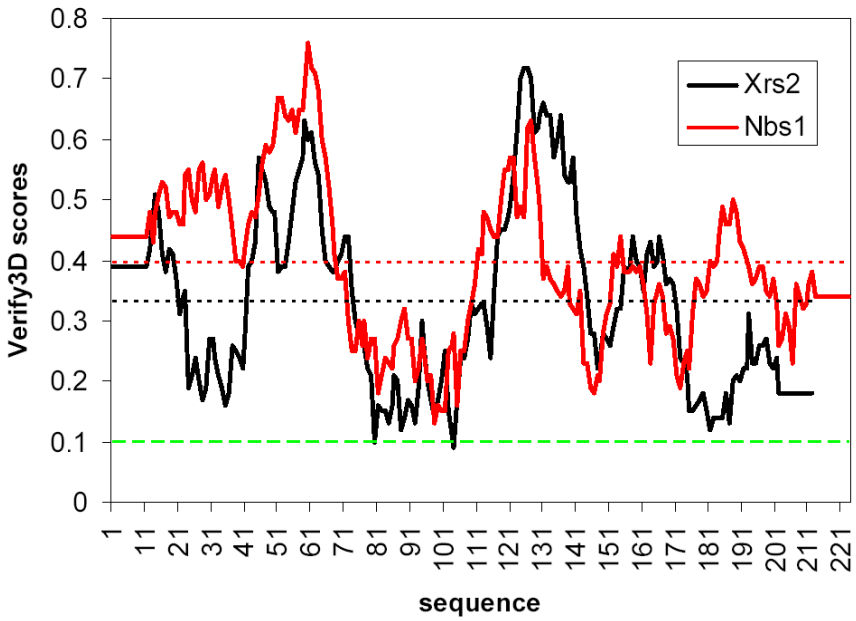


figure 24 : Evaluation des deux modèles de tandem BRCT de Nbs1 (en rouge) et Xrs2 (en noir) par le potentiel statistique VERIFY3D. Le score résidu par résidu est représenté en ordonnée tandis que la séquence est en abscisse.

2.2 Implications fonctionnelles.

2.2.1 Indices suggérant que le tandem BRCT de Nbs1 reconnaît des phospho-sérines.

Dans deux études publiées en 2003, les auteurs identifient les tandems BRCT comme des modules liant des phospho-sérines *in vitro* (Manke et al., 2003; Yu et al., 2003). Ils on pu montrer que la poche de reconnaissance de la phospho-sérine (pS) induit une signature

spécifique au sein de la séquence. Celle-ci se caractérise par un motif [S/T-G] dans la boucle β_1/α_1 couplé à un motif [S/T-x-K] dans l'hélice α_2 du premier domaine BRCT (Glover et al., 2004). La glycine du premier motif n'a de contacts avec le résidu pS que *via* les atomes de son squelette ; c'est probablement la raison pour laquelle ce résidu est moins contraint que les autres résidus de la signature. Par exemple, des tandems de domaines BRCT tels que celui de 53Bp1, possédant une méthionine à cette position, lient également des motifs pS *in vitro*.

Au sein de l'alignement de séquences que nous proposons pour le tandem BRCT de Nbs1 et Xrs2 (présenté **figure 21**), les régions les plus conservées coïncident avec celles contenant des résidus en interaction directe avec la phospho-sérine au sein des structures de tandem BRCT complexés à des phospho-peptides (Stucki et al., 2005). Pour les séquences des organismes allant de l'homme à *Yarrowia lipolytica*, le motif consensuel [S-C/F] dans la boucle β_1/α_1 ainsi que le motif [S/T-x-K] dans l'hélice α_2 sont strictement conservés, ce qui suggère que ces protéines ont la capacité de lier des phospho-sérines. Pour les organismes plus proches de Xrs2, de *Kluyveromyces yarrowii* à *Solenodon paradoxus*, les positions interagissant avec la phospho-sérine par leur chaîne latérale sont conservées mais ne respectent pas la signature consensuelle. Le motif de la boucle β_1/α_1 est remplacé par un motif [P-P], et celui de l'hélice α_2 par [S-x-R]. Cependant, certains indices laissent penser que ces tandems de domaines BRCT pourraient également lier des motifs pS. Tout d'abord, le domaine BRCT de la ligase III, dont il a été montré qu'il lie des motifs pS *in vitro* (Yu et al., 2003), contient également une proline à la place de la sérine dans la boucle β_1/α_1, comme Xrs2. De plus, dans le cas de la ligase IV qui elle aussi lie des motifs pS *in vitro* (Yu et al., 2003), on trouve une arginine à la place de la lysine du motif consensuel de l'hélice α_2, comme dans Xrs2.

Dans le cadre des tandems de domaines BRCT, les travaux de Mark Glover (*University of Alberta*, Canada) en 2004 ont montré que la position pS+3 était principalement responsable de la sélectivité pour certains motifs encadrant le résidu pS (Glover et al., 2004) Les acides aminés entourant la position pS+3 sont situés dans le sillon à l'interface des deux domaines BRCT. Sur l'alignement présenté **figure 21**, les résidus dont la chaîne latérale est en contact avec la position pS+3 sont indiqués par un cercle vert. On constate sur l'alignement que ces positions sont bien conservées.

2.2.2 Importance fonctionnelle du second BRCT : interaction Nbs1 – Mdm2

La protéine Mdm2, pour *murine double minute* 2, est un régulateur crucial de l'activité du suppresseur de tumeurs p53 dont le gène est muté dans 50% des tumeurs (Toledo and Wahl, 2006; Vassilev, 2007). Une étude de l'équipe de Christine Eischen (*Eppley Institute for Cancer Research*, *University of Nebraska*, USA) a montré en 2005 que Mdm2 interagit également avec le complexe MRN. Cette interaction fait intervenir exclusivement Nbs1 ; aucune interaction n'a en effet été détectée avec les autres membres du complexe MRN (Alt et al., 2005).

Pour aller plus loin dans cette étude, l'équipe de Christine Eischen a tenté d'identifier au sein des domaines structuraux et des domaines de liaison connus de Nbs1 lequel médiait l'interaction avec Mdm2. Ils ont pu établir que le fragment 198-314 de Mdm2 était nécessaire à la liaison à Nbs1, mais qu'aucun des domaines structuraux ou de liaison de Nbs1 alors identifiés n'était suffisant pour établir l'interaction Nbs1-Mdm2. Les domaines testés étaient le domaine FHA N-terminal, le premier domaine BRCT (le second n'étant pas encore identifié), et les deux domaines de liaison à Tel1 et Mre11 de l'extrémité C-terminale. Finalement, la région centrale 221-540 de Nbs1 a été identifiée comme région d'interaction.

Cette région 221-540 comprend le second domaine BRCT mis en évidence au cours de notre étude, qui s'étend de la position 221 à la position 330. En aval de ce domaine BRCT, la portion 330-540 est prédite comme non repliée (Ward et al., 2004). Ces résultats suggèrent donc que l'interaction Mdm2-Nbs1 pourrait être médiée par le second domaine BRCT de Nbs1.

2.2.3 Structure de l'assemblage FHA, tandem BRCT

Etant donné que la connexion séparant le domaine FHA N-terminal du premier domaine BRCT est extrêmement courte puisqu'elle ne comprend que deux résidus, nous nous sommes interrogés sur l'organisation du complexe intramoléculaire FHA – tandem BRCT. Pour tester les différents arrangements possibles, nous avons utilisé le programme HADDOCK en ajoutant une contrainte telle que l'extrémité C-terminale du domaine FHA et l'extrémité N-terminale du tandem BRCT soient distantes d'au plus 2Å (Dominguez et al., 2003).

Les complexes générés convergent vers une même solution, illustrée sur la **figure 25**. En raison de contraintes stériques, la distance entre le site de liaison FHA / phospho-thréonine et le site de liaison tandem BRCT / phospho-sérine est toujours supérieure ou égale à 45Å. Dans l'hypothèse où un même fragment protéique serait reconnu par les deux sites d'interaction, cela signifie que les résidus pS et pT doivent être séparés d'une quinzaine de résidus au minimum. Dans l'hypothèse où le domaine FHA et les tandems de domaines BRCT de Nbs1 auraient une action combinée dans le but d'intégrer différents signaux, cette contrainte stérique pourrait être particulièrement importante.

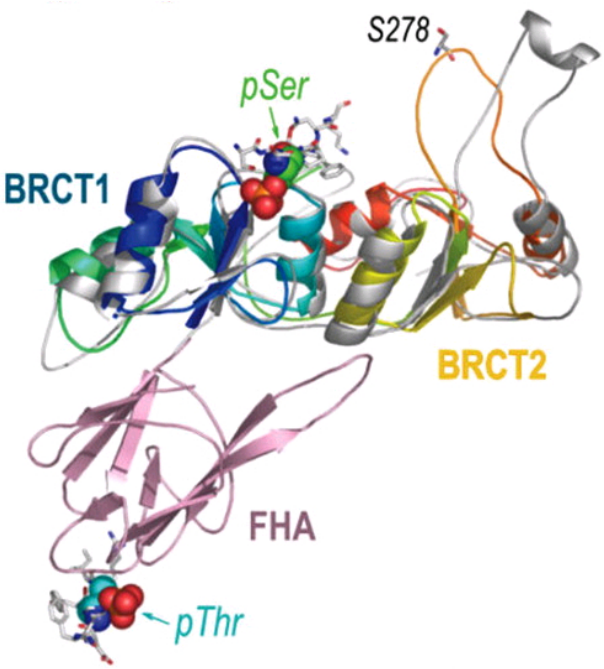


figure 25 : Modèle de l'assemblage intramoléculaire du domaine FHA de Nbs1 avec le tandem BRCT. Le processus d'amarrage moléculaire a été effectué avec un squelette rigide, *via* le logiciel HADDOCK (Dominguez et al., 2003) couplé au programme CNS (Brunger et al., 1998). La structure du tandem BRCT (voire légende **figure 22**), est amarrée au domaine FHA (rose). Les sites de reconnaissance phospho-thréonine du domaine FHA et phospho-sérine du tandem BRCT sont indiqués (phospho-résidus en sphères).

2.3 Perspectives

La mise en évidence de la présence d'un tandem BRCT dans la protéine Nbs1 ouvre de nouvelles hypothèses quant aux mécanismes moléculaires à l'origine du syndrome NBS. De façon remarquable, la mutation présente chez 95% des patients NBS coupe le tandem BRCT au niveau de la connexion séparant les deux domaines. Deux études réalisées par des équipes différentes tendent à montrer que lorsqu'on isole les domaines BRCT d'un tandem,

ceux-ci se replient de façon autonome (Gaiser et al., 2004; Zhang et al., 1998). Nous pouvons donc raisonnablement supposer que chez les patients atteints de NBS, la protéine p26 contiendrait un domaine FHA et un domaine BRCT, tandis que la protéine p70 contiendrait le deuxième BRCT éventuellement replié dans sa région N-terminale avec les sites de fixation à Mre11 et ATM dans sa région C-terminale.

Il a été montré que le domaine FHA et le premier domaine BRCT de Nbs1 étaient capables de se lier *in vitro* aux histones H2AX phosphorylées par le complexe ATM (Kobayashi et al., 2004). La phosphorylation de la sérine 129 est l'un des premiers marqueurs de l'activation des voies de réparation des cassures double-brin (Lowndes and Toh, 2005). Nos résultats suggèrent que chez les patients atteints de NBS, la protéine p26 pourrait se lier aux histones H2AX au niveau de la sérine 129 mais avec une sélectivité moins forte due à l'absence du second domaine BRCT. Cette hypothèse s'appuie sur le fait que dans les tandems BRCT, la position pS+3, qui d'après les travaux de Mark Glover est la principale responsable de la sélectivité pour certains motifs encadrant le résidu pS, interagit avec les résidus du sillon situé à l'interface des deux domaines BRCT (Glover et al., 2004).

Nous avons initié une collaboration avec Noël Lowndes (*National University of Galway*, Irlande), qui étudie le rôle du complexe MRN et celui des protéines de la « famille Nbs1 » chez les eucaryotes supérieurs (van den Bosch et al., 2003). Noël Lowndes dispose actuellement de cellules animales DT40 *NBS1* $^{-/-}$ qui constituent un système intéressant dans lequel on peut envisager d'étudier des mutants de Nbs1 qui ne lierait plus la phospho-sérine 129 des histones H2AX.

Pour la protéine de levure Xrs2, nous ne possédons pas d'évidences permettant de conclure que le tandem BRCT nouvellement détecté en aval du domaine FHA reconnaît des motifs contenant des phospho-sérines. A l'heure actuelle, les cristallographes de notre laboratoire travaillent sur la purification du tandem de domaines BRCT de Xrs2 et du module complet FHA – tandem BRCT, dans l'optique de déterminer la structure de ces modules. Nous espérons que ces études valideront et affineront les modèles structuraux que nous avons proposés. Les propriétés du tandem de domaines BRCT de Xrs2 en tant que module de liaison à des motifs contenant des pS pourront également être étudiées par des approches biochimiques.

Chapitre 3 : Le problème de l'alignement des séquences en vue de la modélisation structurale.

La recherche de l'alignement optimal dans l'optique d'une modélisation comparative est souvent problématique dans le cas des PRMs hautement divergents, comme en témoigne l'exemple traité précédemment avec Nbs1 et Xrs2. Au cours de ce chapitre, nous proposons d'implémenter au sein du programme HMMer une généralisation de l'algorithme de Viterbi. Cette nouvelle fonction, HmmKalign, a pour objectif d'explorer l'espace des alignements de séquences en se focalisant sur les alignements les plus probables.

3.1

Introduction.

Dans cette section, nous présentons les développements méthodologiques effectués pour générer automatiquement un ensemble d'alignements sous-optimaux plutôt qu'un unique alignement optimal (nommé OSA pour *Optimal Sequence Alignment)*. Plusieurs aspects intéressants de l'exploration ciblée des alignements peuvent être soulignés.

Dans le cadre des PRMs hautement divergents l'alignement correct est rarement l'OSA. La question se pose alors de savoir si l'alignement correct peut être détecté, au moins en partie, dans le voisinage de l'OSA. Ensuite, l'exemple des domaines BRCT de Nbs1 montre que l'alignement a du être optimisé de façon itérative pour atteindre un modèle satisfaisant. Générer de manière automatique un ensemble d'alignements alternatifs dont les scores probabilistes soient élevés serait une façon de faciliter voire de remplacer ces ajustements manuels.

Dans cette optique, nous avons recherché une méthode qui permette d'explorer le voisinage de l'alignement optimal (optimal pour ce qui est de l'espace des alignements de séquences). Comme mentionné dans l'introduction, le formalisme des modèles de Markov cachés offre un cadre particulièrement efficace pour traiter les alignements de séquences divergentes et a donc été sélectionné pour atteindre notre objectif. Pour produire des alignements alternatifs dans le voisinage de l'alignement des séquences optimal, trois solutions ont été envisagées.

(i) La première consiste à introduire des perturbations aléatoires dans l'OSA. Cependant, ces perturbations sont susceptibles de dégrader significativement la qualité des alignements et ne facilitent pas forcément la sélection d'un alignement alternatif intéressant.

Les alternatives consistent à modifier directement l'algorithme de Viterbi utilisé pour détecter l'OSA (Forney, 1973; Viterbi, 1967). Deux variantes de cet algorithme permettent d'accéder à l'ensemble des alignements dans un voisinage fixe de l'OSA.

(ii) On peut fixer le voisinage en terme de distance ε par rapport à l'OSA et chercher à générer tous les alignements dont le score est compris entre le meilleur score et

le meilleur score +ε, ε correspondant la distance par rapport à l'OSA. Cela suppose de remplacer l'algorithme de Viterbi par une adaptation de l'algorithme présenté par Waterman et Byers dans le cadre des alignements entre paires de séquences (Waterman and Byers, 1985). Le désavantage de cette approche est que le nombre d'alignements générés n'est pas estimable *a priori* et peut s'avérer très important : un problème de gestion de l'espace mémoire peut se poser. De plus, si le critère de discrimination entre alignements est basé sur la production de modèles structuraux et leur évaluation par des potentiels statistiques (démarche identique à celle mise en place pour Nbs1), cette procédure peut s'avérer très lente.

(iii) On peut également fixer le voisinage de l'OSA en terme de nombre κ d'alignements ; ce qui signifie que les κ meilleurs alignements sont calculés. Ceci est possible en substituant l'algorithme de Viterbi par sa variante N-Viterbi utilisée pour résoudre des problèmes de reconnaissance de parole (Huang and Chiang, 2005). Il présente l'avantage de contrôler le nombre de modèles à évaluer.

L'examen de ces différentes stratégies nous a conduit à implémenter la troisième option, l'algorithme de N-Viterbi, qui permet de récupérer les κ alignements dont les scores sont les plus élevés (contrairement à la méthode des perturbations) tout en maîtrisant le nombre d'alignements produits.

3.2 Exploration ciblée de l'espace des alignements séquence-HMM au voisinage de l'alignement optimal.

3.2.1 Implémentation de la fonction HMMKALIGN au sein de HMMER.

Formellement, aligner une séquence $s_{obs} = s_1 \ldots s_T$ sur un modèle de Markov caché consiste à identifier la séquence d'états qui maximise la probabilité d'émission de s_{obs}. Comme nous l'avons introduit précédemment, l'algorithme utilisé pour trouver cette séquence est l'algorithme de Viterbi, dont la complexité en temps est de $O(TM)$, lorsque T est la longueur de s_{obs} et M le nombre d'états contenus dans le modèle de Markov caché (Forney, 1973; Viterbi, 1967).

Pour déterminer l'ensemble des κ séquences d'états qui maximisent l'émission de s_{obs}, il est possible d'utiliser une généralisation de l'algorithme de Viterbi dont nous présentons ici une version naïve.

Procédure Viterbi_Généralisé (*HMM*, κ, s_{obs})

pour tous les acides aminés *aa* de s_{obs} **faire** :
    /* parcours des sommets du HMM dans l'ordre topologique */
    pour tous les sommets *v* du *HMM* **faire** :
    l_v = {}
        pour toutes les arêtes *e={u,v}* incidentes à *v* **faire** :
            /* l_u contient les κ meilleurs scores associés au parcours du *HMM* finissant en *u* */
            /* *w(e)* est le poids de l'arête *e* (ou probabilité de transition de *u* vers *v*) */
            pour tous les scores *score* de l_u **faire** :
                score ← *score* * *w(e)*
                l_v ← concaténer (l_v, *score*)
            fin
        /* les κ meilleurs scores de l_v sont conservés et triés */
        l_v ← trier (l_v)
        l_v ← $l_v[1 : \kappa]$
        fin
    fin
/* le vecteur l_v de l'état final contient les κ meilleurs scores associés aux κ meilleurs parcours du *HMM* */
fin

algorithme 1 : Algorithme de Viterbi généralisé (version naïve, les conditions initiales et finales ne sont pas traitées), permettant de rechercher pour une séquence s_{obs} donnée, les κ parcours d'un HMM aboutissant aux κ plus fortes probabilités d'émission.

Cet algorithme a une complexité en temps de $O(TM\kappa log\kappa)$. Par rapport à la version classique de l'algorithme de Viterbi, le facteur multiplicatif $\kappa log\kappa$ provient des tris nécessaires pour sélectionner les κ meilleures solutions au fur et à mesure du parcours du modèle de Markov caché.

Nous avons implémenté une version améliorée de cet algorithme de Viterbi généralisé au sein du programme HMMer, dont les codes sources sont accessibles et distribués librement (Eddy, 1996). Dans cette version améliorée, l'utilisation de listes de priorité à la place de vecteurs de scores permet de diviser le temps d'exécution du programme par une constante égale au nombre moyen d'arêtes par noeud de l'automate (Huang and Chiang, 2005). Les détails de l'implémentation sont donnés en annexe (**Annexe A**).

3.2.2 Influence des méthodes de construction du HMM.

L'architecture traditionnelle des modèles de Markov cachés utilisés pour les alignements de séquences protéiques suit soit le Plan9 soit le Plan7 (**figure 15**). Ces deux plans, utilisés respectivement dans SAM (Karplus et al., 1998; Karplus et al., 2005) et HMMer (Eddy, 1996), considèrent 3 états par nœud :

- l'appariement (*match state*, M_i) ;
- l'insertion (*insertion state*, I_i) ;
- ou la délétion (*deletion state*, D_i).

Le fait de considérer une colonne de l'alignement multiple servant à paramétrer le HMM comme un état d'appariement ou comme une insertion entre deux états d'appariement consécutifs est un choix crucial dans le contexte de la génération d'alignements sous-optimaux.

L'importance de choisir judicieusement quelles colonnes de l'alignement multiple doivent être considérées comme des appariements est illustrée par la **figure 26**. La partie **(A)** montre l'alignement de départ servant à construire et paramétrer le HMM.

La méthode traditionnelle, utilisée par HMMer et SAM, (**figure 26**-B) utilise la conservation et le nombre de « gaps » dans chaque colonne de l'alignement multiple comme critères pour définir l'existence d'un état d'appariement dans ces colonnes (en gris). Dans ce cas, des régions de boucles variables reliant deux structures secondaires conservées peuvent correspondre à des états d'appariement. La génération d'alignements sous-optimaux sur un tel HMM produira principalement des variations focalisées dans ces régions de boucles (**figure 26**-C). Or, l'alignement et la modélisation de séquences divergentes pose surtout des problèmes dans le positionnement correct des régions des structures secondaires. Les boucles sont généralement modélisées dans une procédure spécifique au cours de laquelle l'alignement n'intervient pas. Il est donc probable que la sélection des états d'appariements par la méthode traditionnelle produira des alignements sous-optimaux qui n'améliorent la qualité des modèles que de façon marginale.

Afin de pallier à cette limitation, nous avons créé une option dans HMMKALIGN qui permet de modifier les règles traditionnelles de sélection des états d'appariements. Avec cette option, seules les régions de structure secondaires conservées sont considérées comme des états d'appariements (**figure 26**-D). Comme l'illustre le même exemple que précédemment, une telle contrainte permet de générer des alignements sous-optimaux qui diffèrent au sein des structures secondaires conservées (**figure 26**-E). Par la suite, nous avons évalué les performances des modèles de Markov caché « contraints » par les positions correspondant à des structures secondaires conservées et nous les avons comparé aux alignements sous-optimaux générés lorsque le modèle est construit de manière « traditionnelle ».

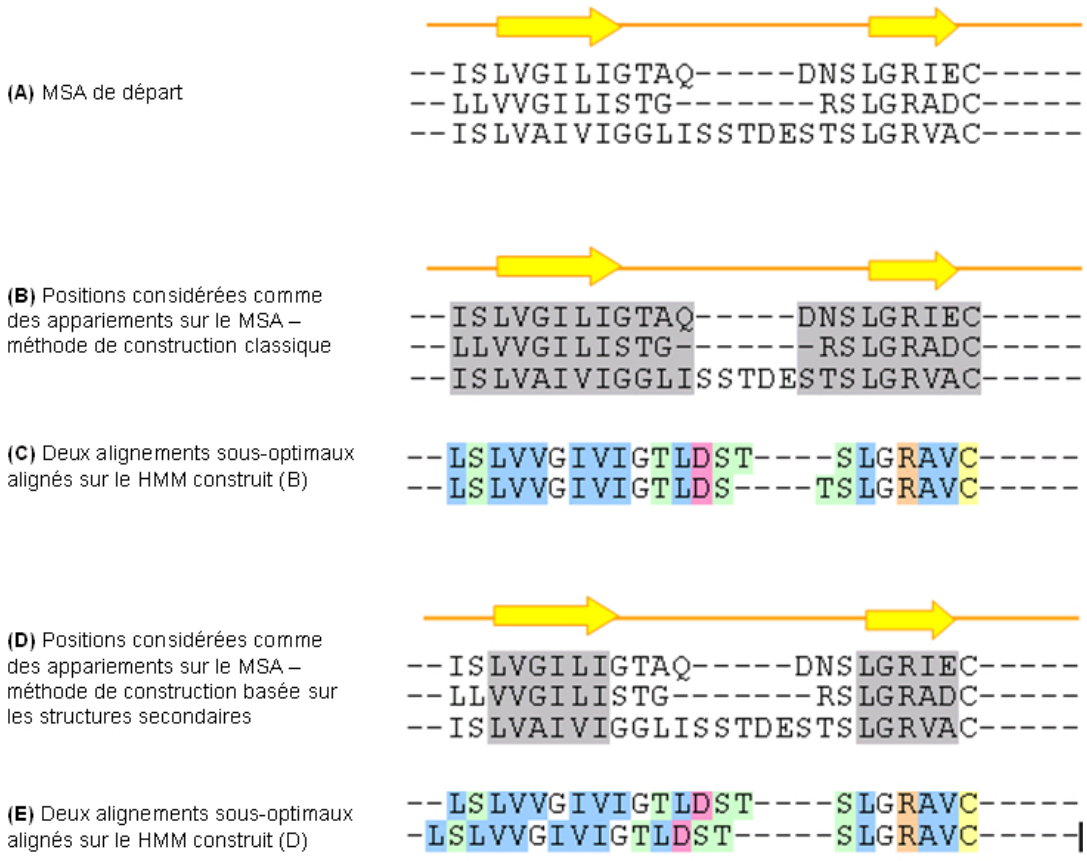


figure 26 : Importance du choix de l'architecture du HMM. **(A)** Alignement multiple de départ, duquel doivent être déduits les paramètres du HMM. La ligne du haut représente les structures secondaires conservées (ici deux brins β). **(B)** Méthode de construction du HMM traditionnelle ; les positions qui seront considérées comme des appariements sont surlignées en gris. **(C)** Les deux séquences représentent deux alignements sous-optimaux générés en alignant une même séquence sur le HMM construit de façon traditionnelle. **(D)** Méthode de construction du HMM où les états d'appariements sont contraints par la position des structures secondaires ; les positions qui seront considérées comme des appariements sont surlignées en gris, elles correspondent parfaitement aux régions de structures secondaires conservées. **(E)** Les deux séquences représentent deux alignements sous-optimaux générés an alignant une même séquence sur le HMM dont les états d'appariements sont contraints par le positionnement des structures secondaires.

3.2.3 Base d'alignements tests de familles de séquences divergentes.

Pour construire la base de test, nous avons utilisé les alignements structuraux provenant de la banque de données HOMSTRAD (Mizuguchi et al., 1998; Stebbings and Mizuguchi, 2004). Nous avons sélectionné 23 familles qui respectaient les critères suivants :

(i) l'alignement structural comporte au minimum 5 structures ;

(ii) l'identité de séquence moyenne au sein de la famille est inférieure ou égale à 25%.

La liste complète des familles retenues est présentée en **annexe A**. Au sein de chaque famille, 5 séquences s_{obs} ont été sélectionnées aléatoirement. Le processus de test se déroule ensuite de la façon suivante : (1) la séquence s_{obs} est extraite de l'alignement structural ; (2) le profil de la famille est construit à partir des séquences restantes à l'exception de celles qui présentent plus de 40% d'identité de séquence avec s_{obs} (pour ne pas introduire de biais) ; (3) on aligne s_{obs} sur le HMM construit ; (4) on compare cet alignement avec l'alignement structural de départ. Le fait de choisir un nombre identique de séquences dans chaque famille permet de ne pas sur-représenter les familles les plus abondantes.

3.2.4 Mesures utilisées pour évaluer la qualité des alignements.

Pour comparer un alignement prédit à l'alignement correct (ici l'alignement structural), les scores traditionnellement utilisés sont le Q_{mod}, le Q_{dev} et le Q_{local} (Yona and Levitt, 2002).

En considérant que l'alignement structural possède une longueur L_s, l'alignement prédit possède une longueur L_p, et que dans l'alignement prédit N positions sont correctement alignées, on a :

$$Q_{mod} = N / L_p \qquad \text{(formule 1)}$$

$$Q_{dev} = N / L_s \qquad \text{(formule 2)}$$

Intuitivement, le Q_{mod} permet de calculer la fraction de positions de l'alignement prédit correctement alignées relativement au nombre total de positions dans l'alignement prédit. Le Q_{dev} ramène cette fraction au nombre global de positions dans l'alignement structural.

Avec ces deux mesures, les décalages au sein des alignements sont considérés comme des positions mal alignées, sans que l'amplitude de ces décalages ne soit prise en compte. Le Q_{local} a été introduit pour traiter ce problème spécifique. Son expression est la suivante :

$$Q_{local} = \sum_{x=1}^{Lp} \frac{(1/2)^{s(x)}}{Lp} \qquad \text{(formule 3)}$$

où $s(x)$ représente le décalage entre la position du résidu x dans l'alignement prédit et sa position dans l'alignement structural. De cette façon, pour chaque résidu correctement aligné, on a $s(x) = 0$ et le Q_{local} augmente de $0.5^{0} / L_p = 1/L_p$. Pour les résidus mal alignés, seuls les très petits décalages génèreront une modification significative du Q_{local}.

3.2.5 Mesure utilisée pour évaluer la diversité des alignements.

Pour évaluer la divergence des alignements sous-optimaux générés par HMMKALIGN, nous utilisons l'alignement multiple des différents alignements sous-optimaux générés. A partir de cet alignement de κ séquences, nous pouvons calculer l'entropie de Shannon H (Shannon, 1948) de toutes les positions p le long de l'alignement :

$$H(p) = -\sum_{i \in A} P(p,i) \log_2 P(p,i) \quad \text{(formule 4)}$$

où A est l'ensemble des 20 acides aminés, $P(p,i)$ est le nombre d'occurrences de l'acide aminé i à la position p divisé par le nombre κ de séquences.

3.2.6 Procédure de test.

Notre base de test comporte 23 familles hautement divergentes au sein desquelles cinq séquences sont isolées puis réalignées en aveugle. On dispose donc de 23*5=115 alignements test. Pour ces 115 alignements test, nous avons utilisé la fonction HMMKALIGN pour calculer 20 alignements sous-optimaux pour chacun des deux modes de construction du HMM décrits en section 3.2.2, l'option de construction « traditionnelle » et l'option de construction spécifiquement « contrainte » par les structures secondaires conservées.

3.3 Résultats obtenus par HmmKalign sur 115 alignements test (κ=20).

3.3.1 Diversité au sein des 20 alignements sous-optimaux générés.

L'idée de contraindre les états d'appariements par les régions de structure secondaire a été introduite dans le but de générer des alignements plus hétérogènes. Dans ce paragraphe, nous allons tester si cette propriété est bien vérifiée en utilisant l'entropie de Shannon (voir 3.2.5) comme quantificateur de divergence.

La **table 5** reporte les résultats obtenus. Sur la séquence entière, en générant 20 alignements sous-optimaux avec la méthode de construction du modèle traditionnelle, l'entropie a une valeur moyenne de 0.17 et atteint 0.87 au maximum. Lorsque la méthode de construction contraignant les états d'appariement est utilisée, cette entropie a une valeur moyenne de 0.25 et atteint au maximum 0.93. Afin d'évaluer la significativité de ces différences d'entropie, un test de Student à été réalisé. Dans le cadre des grands nombres ($n \geq 30$) ce test est robuste et applicable aux échantillons non gaussiens. L'hypothèse H0 testée est la suivante : $m_1 = m_2$; où m_1 et m_2 sont les moyennes des deux distributions. Sur la totalité de la séquence, l'hypothèse H0 est rejetée avec un risque de 0.1% d'erreurs. Restreinte aux régions de structures secondaires, elle est rejetée avec un risque de 5% d'erreurs (détails en annexe B). En conclusion, les différences observées entre les moyennes sont donc significatives.

Les figures suivantes permettent de visualiser la distribution de l'entropie, soit sur la totalité de l'alignement (**figure 27**-A), soit uniquement sur les régions correspondant aux structures secondaires (**figure 27**-B). On y constate que les distributions sont sensiblement différentes dans les deux cas (conformément au test de Student) et que l'exploration de l'espace des alignements de séquence est plus importante avec la méthode de construction des HMM « contrainte ».

	Traditionnal building method (sequence conservation)	Alternative building method (second. struct. conservation)
Entropy over the whole alignments		
mean	0.1698	0.2539
standard dev	0.0166	0.0205
max	0.8750	0.9335
Entropy over the secondary structure elementsof the alignments		
mean	0.1239	0.1617
standard dev	0.0179	0.0179
max	0.9514	0.9695

table 5: Comparaison de l'entropie moyenne au sein des 20 alignements sous-optimaux calculés, en fonction de la méthode utilisée pour construire le modèle de Markov caché : la méthode classique, basée sur la conservation de séquence (colonne du milieu), et la méthode basée sur la conservation des structures secondaires (colonne de droite). On différencie les résultats sur l'ensemble de la séquence, et uniquement sur les portions faisant partie de structures secondaires. Les résultats présentés sont la moyenne des 115 entropies, l'écart type des 115 entropies et le maximum des 115 entropies.

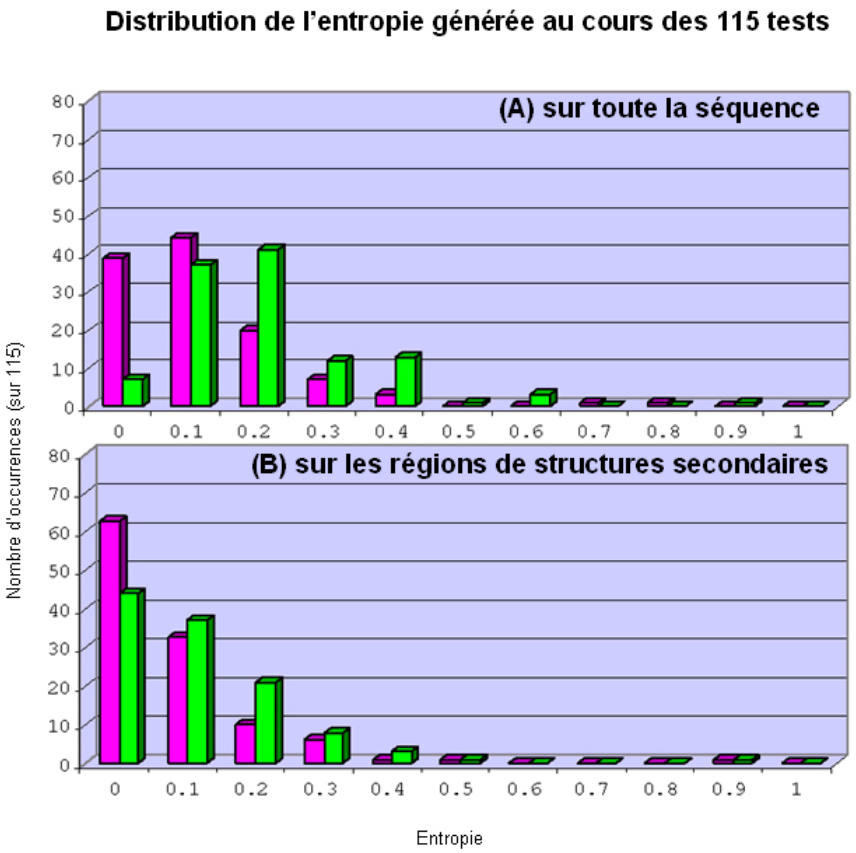


figure 27 : **(A)** Entropie sur l'ensemble de la séquence test. En rose, la distribution de l'entropie au sein des 20 alignements sous-optimaux lorsque le modèle de Markov caché est construit de manière traditionnelle. En vert, cette même distribution lorsque le modèle de Markov caché est construit en se basant sur la conservation de structures secondaires. **(B)** Idem, mais en restreignant les calculs d'entropie aux seules régions de structures secondaires.

3.3.2 Amplitude des améliorations obtenues pour les Q_{mod}, Q_{dev}, et Q_{local}.

Dans ce paragraphe, la qualité des alignements sous-optimaux générés en terme de Q_{mod}, Q_{dev} et Q_{local} est évaluée. Nous avons généré 20 alignements sous-optimaux avec chacune des deux méthodes de construction du modèle de Markov caché, et nous souhaitons savoir si au sein de ces alignements, il existe un « meilleur » alignement que l'OSA (*Optimal Sequence Alignment*) traditionnellement généré par HMMER.

Pour évaluer le potentiel d'amélioration de qualité des alignements, nous proposons de calculer la différence entre le score $Q_i(p)$ du meilleur alignement sous-optimal p et le score $Q_i(OSA)$ de l'OSA :

$$\Delta Q_i = \max_{p=1\ldots20} \left(Qi(p)\right) - Qi(OSA) \qquad \text{(formule 6)}$$

Plus cette différence ΔQ_i est importante, plus l'amélioration est significative. A titre d'exemple, si le $\Delta Q_i = 0.05$ pour le Q_{mod}, cela signifie qu'il existe un alignement sous-optimal dont le Q_{mod} est supérieur de 0.05 à celui de l'OSA. On peut ainsi évaluer le nombre d'alignements améliorés de façon à ce que l'amplitude de l'amélioration soit supérieure à une valeur seuil s donnée.

Les améliorations obtenues par l'exploration des 20 alignements sous-optimaux sont présentées au sein de la **table 6** et de la **table 7**. Lorsque le modèle de Markov caché est construit traditionnellement (**table 6**), les améliorations sont très fréquentes mais leur amplitude est souvent faible. En effet, pour le Q_{mod}, alors que plus de 70% des alignements sont améliorés (première ligne, $\Delta Qi > 0.00$), seuls 10% des alignements ont un Q_{mod} qui augmente de plus de 0.10 (dernière ligne, $\Delta Qi > 0.10$). Cette propriété est également vérifiée pour le Q_{dev} et le Q_{local}. A l'inverse, lorsque le modèle de Markov caché initial est construit en contraignant les états d'appariement aux régions des structures secondaires (**table 7**), les améliorations sont moins nombreuses mais leur amplitude est plus souvent importante. Pour le Q_{mod}, la variation est supérieure à 0.10 pour 17% des alignements contre 10% précédemment. Avec le même seuil, pour le Q_{dev} et le Q_{local}, le nombre d'alignements améliorés passe respectivement de 9% à 22% et de 3% à 14%.

Ces résultats montrent que les améliorations les plus importantes sont obtenues lorsque la construction du modèle restreint les états d'appariement aux régions de structures secondaires. L'ensemble des résultats obtenus jusqu'à présent suggère donc que les alignements sous-optimaux produits lorsque les états d'appariement du HMM sont contraints par la position des structures secondaires sont (i) plus divergents les uns des autres et de ce fait (ii) permettent d'obtenir des améliorations plus conséquentes de la qualité des alignements.

ΔQ_i	Q_{mod}	Q_{dev}	Q_{local}
0.00	83/115 (72%)	80/115 (70%)	87/115 (76%)
0.01	79/115 (69%)	78/115 (68%)	85/115 (74%)
0.02	53/115 (46%)	57/115 (50%)	50/115 (44%)
0.05	23/115 (20%)	25/115 (22%)	17/115 (15%)
0.10	11/115 (10%)	10/115 (9%)	4/115 (3%)

table 6 : Comparaison du Q_{mod}, Q_{dev} et Q_{local} de l'OSA et des alignements sous-optimaux lorsque l'architecture du modèle de Markov caché est construite traditionnellement. La première colonne indique le seuil *s* utilisé (seuil strict : $\Delta Q_i > s$). Dans la seconde colonne, on dénombre les alignements où ΔQ_i du Q_{mod} est supérieur ou égal au seuil *s*. Les résultats sont donnés en nombre d'occurrences (sur 115) et en pourcentage de l'ensemble de la base de test. Les colonnes suivantes indiquent les mêmes résultats respectivement pour le Q_{dev} et le Q_{local}.

ΔQ_i	Q_{mod}	Q_{dev}	Q_{local}
0.00	61/115 (53%)	68/115 (59%)	59/115 (51%)
0.01	60/115 (52%)	68/115 (59%)	59/115 (51%)
0.02	49/115 (42%)	55/115 (48%)	45/115 (39%)
0.05	34/115 (30%)	46/115 (40%)	30/115 (26%)
0.10	20/115 (17%)	25/115 (22%)	16/115 (14%)

table 7: Comparaison du Q_{mod}, Q_{dev} et Q_{local} de l'OSA et des alignements sous-optimaux lorsque l'architecture du modèle de Markov caché est retreinte aux régions dont la structure secondaire est conservée. Légende identique à celle de la **table 6**.

Cependant, nous n'avons pas encore évalué si les deux approches amélioraient les mêmes alignements, et dans les mêmes proportions. Afin d'évaluer le recouvrement entre les deux approches, nous avons comparé pour les trois scores Q_{mod}, Q_{dev} et Q_{local} :

- le nombre d'alignements améliorés uniquement par l'exploration des 20 alignements sous-optimaux lorsque le HMM est construit de manière traditionnelle ;
- le nombre d'alignements améliorés uniquement par l'exploration des 20 alignements sous-optimaux lorsque le HMM est construit en contraignant les états d'appariement aux régions des structures secondaires conservées ;
- le nombre d'alignements améliorés par les deux approches.

Les résultats sont présentés dans les trois graphes de la **figure 28**.

Du point de vue du Q_{mod}, l'utilisation conjointe des deux approches permet d'améliorer 95 alignements sur 115, dont 26 pour lesquels l'amélioration peut dépasser le seuil de 0.10. Sur ces 26 alignements, la méthode de construction « contrainte » des HMM par la conservation des structures secondaires est la plus avantageuse car elle concerne 20 alignements dont 15 de façon exclusive. Le recouvrement entre les deux méthodes est fort lorsqu'on s'intéresse au nombre d'améliorations sans prendre en compte leur amplitude (plus de 50% de recouvrement lorsque le seuil $s > 0.00$) et devient faible lorsqu'on s'intéresse aux améliorations importantes (moins de 20% de recouvrement lorsque le seuil $s > 0.10$).

Les résultats pour le Q_{dev} et le Q_{local} suivent la même logique que ce qui a été observé pour le Q_{mod} : alors que pour les faibles améliorations, le recouvrement entre les deux approches est important, celui-ci devient faible lorsqu'on considère uniquement les améliorations plus conséquentes.

Au regard de ces résultats, les deux méthodes sont donc complémentaires et il semble judicieux de les utiliser simultanément pour maximiser l'exploration de l'espace des alignements de séquence.

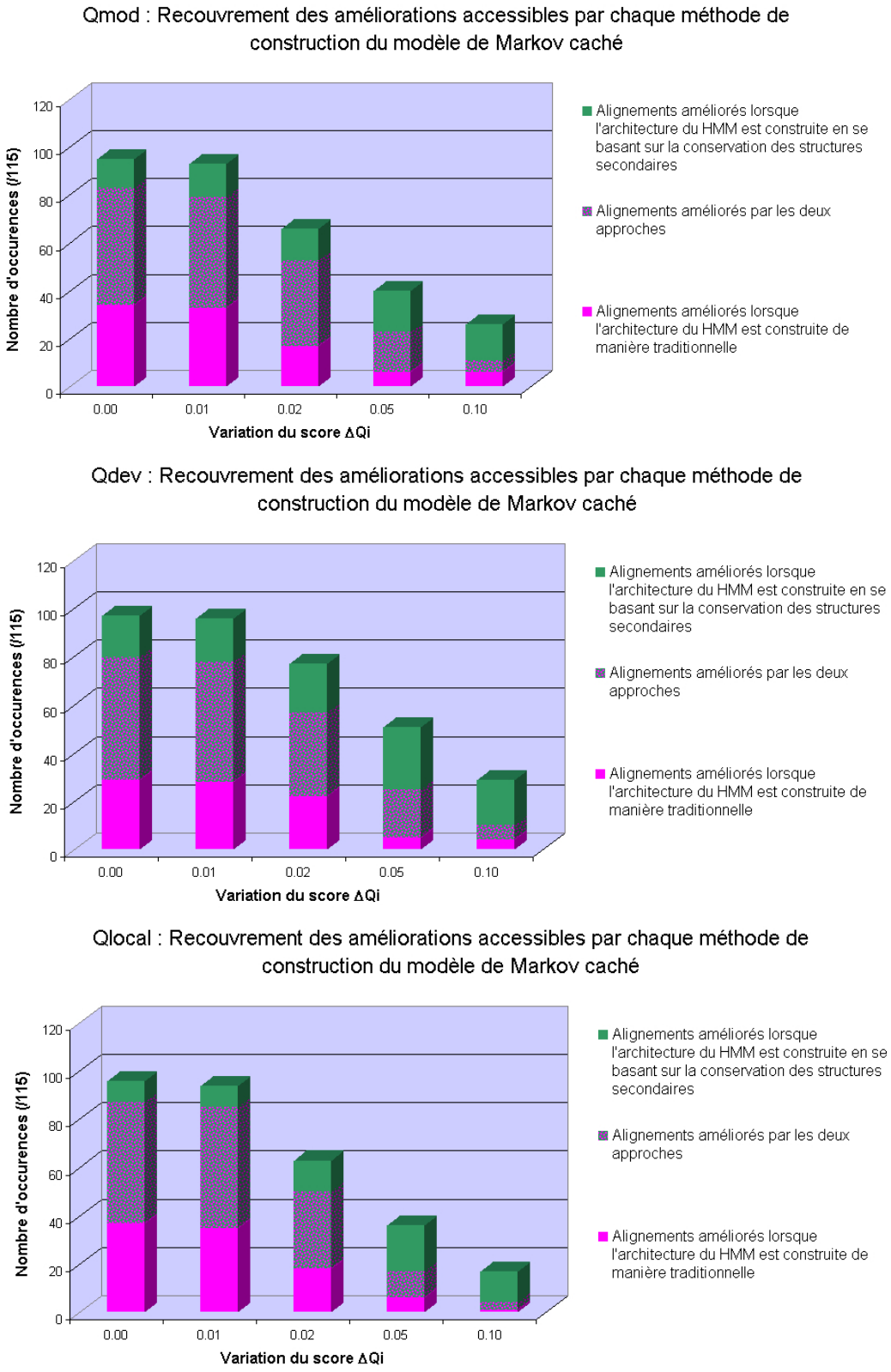


figure 28 : Recouvrement des alignements améliorés. Dans les trois graphes on compare le nombre d'améliorations supérieures à un certain seuil ΔQi que l'on peut obtenir en explorant (i) les 20 alignements sous-optimaux générés lorsque le HMM est construit de manière traditionnelle ; (ii) les 20 alignements sous-optimaux générés en restreignant les états d'appariements aux régions des structures secondaires. Les alignements améliorés uniquement par l'approche (i) sont en rose ; ceux améliorés uniquement par l'approche (ii) sont en vert et les alignements améliorés par (i) et (ii) sont en rose et vert.

3.3.3 Comparaison des moyennes et écart-type du Q_{mod}, Q_{dev} et Q_{local}.

Sur les 115 alignements de la base de test issus de familles de séquences divergentes, le Q_{mod} moyen de l'OSA vaut 0.78 (voir **table 8**). Lorsqu'on explore le voisinage cet alignement, en combinant les 20 alignements sous-optimaux générés obtenus avec chacune des méthodes de construction des HMM, le Q_{mod} moyen atteint 0.84 (**table 8**). On constate une augmentation du même ordre de grandeur pour le Q_{dev} et le Q_{local}. Ces différences entre moyennes sont significatives avec un risque d'erreur inférieur à 1% pour le Q_{mod} et le Q_{dev}, et inférieur à 5% pour le Q_{local} (test de Student).

Ces moyennes cachent cependant des disparités que l'on peut visualiser sur les distributions présentées dans la **figure 29**. Les résultats montrent par exemple que le nombre d'alignements exacts du point du vue du Q_{mod} (Q_{mod}=1) est multiplié par 10 lorsque l'on explore le voisinage de l'alignement optimal (**figure 29**-A). A l'inverse, le nombre d'alignements dont le Q_{mod} est inférieur à 0.5 est divisé par 4.

	Qmod	Qdev	Qlocal
Moyenne et Ecart-type des 115 OSA			
moyenne	0.7848	0.7635	0.8244
écart type	0.1819	0.1902	0.1687
Moyenne et Ecart-type sur le meilleur alignement faisant partie des 20 alignements sous-optimaux (méthode de construction du HMM traditionnelle)			
moyenne	0.8183	0.7986	0.8529
écart type	0.1739	0.1829	0.1635
Moyenne et Ecart-type sur le meilleur alignement faisant partie des 20 alignements sous-optimaux (méthode de construction du HMM basée sur les structures secondaires)			
moyenne	0.8100	0.8040	0.8487
écart type	0.1576	0.1634	0.1408
Moyenne et Ecart-type sur le meilleur alignement faisant partie des 40 alignements sous-optimaux (utilisation conjointe des deux méthodes)			
moyenne	0.8427	0.8331	0.8760
écart type	0.1509	0.1582	0.1375

table 8 : Scores moyens et écarts types du Q_{mod}, Q_{dev} et Q_{local}. On compare (1) les 115 OSA classiquement générés par HMMer ; (2) les 115 meilleurs alignements existant dans le voisinage de l'OSA en explorant les 20 alignements sous-optimaux générés sur le HMM classique ; (3) les 115 meilleurs alignements existant dans le voisinage de l'OSA lorsqu'on explore les 20 alignements sous-optimaux générés sur le HMM dont la construction est basée sur la conservation des éléments de structures secondaires ; (4) les 115 meilleurs alignements existants dans le voisinage de l'OSA en explorant 40 alignements sous-optimaux (utilisation conjointe des deux approches de construction du HMM).

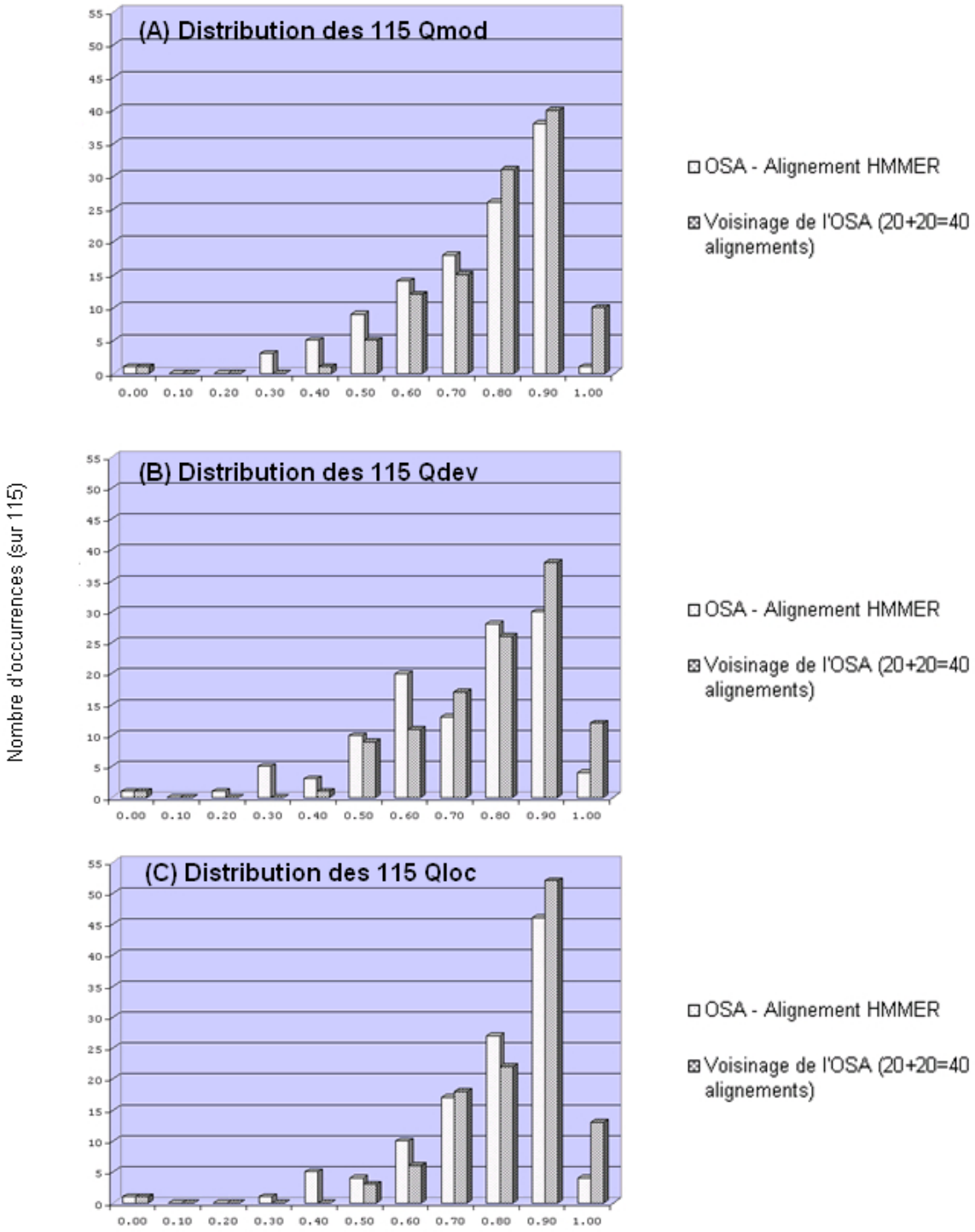


figure 29 : Distribution des 115 Q_{mod}, Q_{dev} et Q_{local} obtenus en considérant l'alignement de séquence optimal (barres claires) et en prenant en compte le voisinage de cet alignement (barres foncées). Afin de maximiser l'exploration de l'espace des alignements, 20 alignements ont été générés avec hmmkalign, k=20, lorsque le HMM est construit de manière traditionnelle et 20 alignements ont étés générés avec hmmkalign, k=20, lorsque le HMM est basé sur les structures secondaires conservées de la famille.

3.3.4 Etude d'un exemple au sein de la famille des thioredoxines.

La famille des thioredoxines rassemble de petites enzymes participant à des réactions d'oxydoréduction. Ces domaines comptent environ 100 acides aminés et leurs séquences sont très divergentes (entre 15 et 20% d'identité de séquence au sein de la famille). Le repliement sous forme de *3-layers sandwich* associé à la famille est bien conservé.

Parmi les membres de la famille des thioredoxines référencés dans la banque de données HOMSTRAD, on trouve la structure de la forme oxydée de la glutarédoxine du bactériophage T4 résolue par diffraction des rayons X en 1992 (Eklund et al., 1992). Cette structure est présentée sur la **figure 30**.

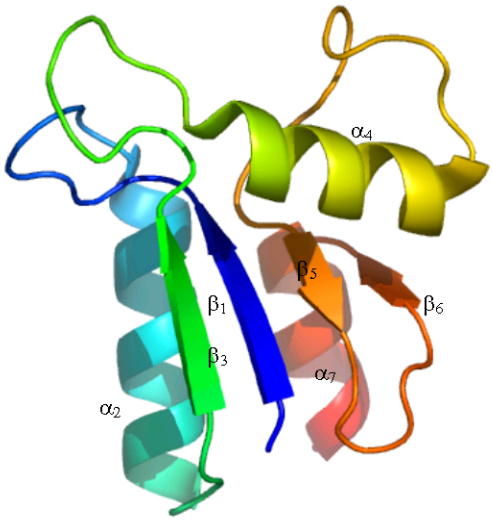


figure 30 : Structure de la forme oxydée de la glutarédoxine du bactériophage T4 (code PDB 1aaz, chaîne A), représentée sous forme de ruban. Un gradient de couleur (du bleu au rouge) indique la position dans la séquence (du N-terminal vers le C-terminal). Les sept structures secondaires sont indiquées.

Aligner la séquence de la forme oxydée de la glutarédoxine du bactériophage T4 avec les séquences des autres membres de la famille est une tâche délicate. L'OSA est clairement loin de l'alignement structural ; son Q_{mod} est égal à 0.50 exactement (figure 32-B, OSA). Plus précisément, les trois premières structures secondaires, le brin β_1, l'hélice α_2 et le brin β_3, ne sont pas reconnus comme faisant partie du domaine thiorédoxine. De plus, l'hélice α_4 est alignée avec les positions correspondant à l'hélice α_2, alors que des insertions sont alignées en face des positions du brin β_3. Au final, seules les trois dernières structures secondaires, les brins β_5 β_6 et l'hélice α_7 sont correctement alignées.

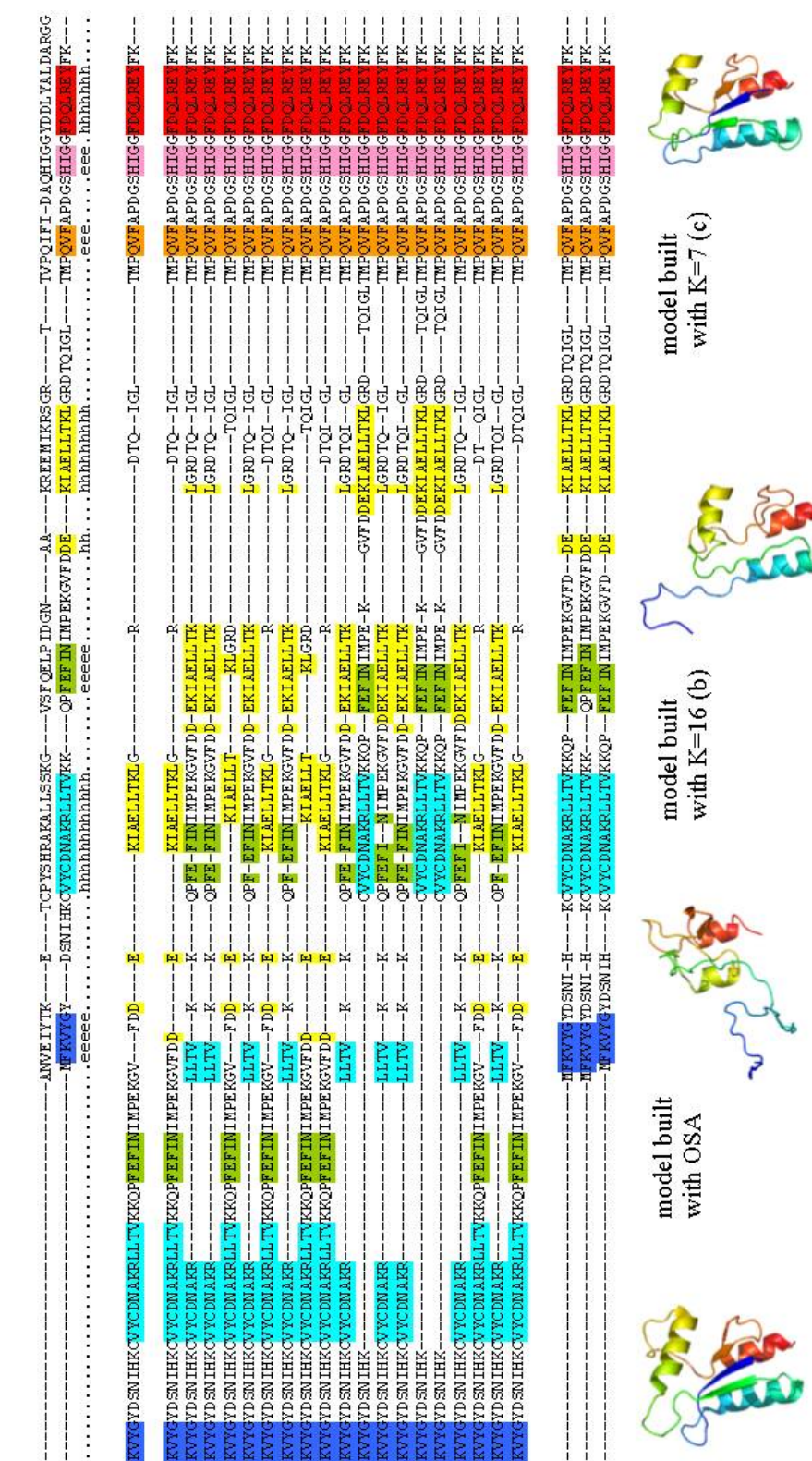


figure 32: Exemple de l'alignement de la forme oxydée de la glutarédoxine du bactériophage T4 (chaîne A, code pdb 1aaz). L'alignement multiple de la séquence avec les autres membres de la famille est représenté à travers une projection en alignement pairwise entre 1aaz et 3grx. Les acides aminés de 1aaz correspondant à une structure secondaire sont surlignés en couleur (le code des couleurs respecte celui de la **Ufigure 30**.(a) Les deux premières lignes présentent l'alignement structural. Les structures secondaires assignées à 1aaz ont été obtenues par les annotations de la banque HOMSTRAD (méthode de calcul JOY (Mizuguchi Deane Bioinformatics 1998)). (b) Les 20 lignes suivantes correspondent aux 20 alignements générés avec la commande hmmkalign lorsque la méthode de construction du modèle de Markov caché est classique. Tous les alignements sont distincts mais il peut arriver que leurs projections sous forme d'alignement pairwise soient identiques (voir par exemple K=12,15 et 16). Au sein des 20 alignements calculés, le premier (K=1) correspond à l'alignement de séquence optimal (OSA). (c) Alignements sous-optimaux générés lorsque l'architecture du modèle de Markov caché est basée sur la conservation des structures secondaires. Les 20 meilleurs alignements ont été calculés mais seuls trois d'entre eux sont présentés (correspondant aux 7eme, 12eme et 19eme meilleurs alignements).

Alignements sous-optimaux avec la méthode traditionnelle de construction du HMM. Nous avons utilisé la fonction HMMKALIGN pour générer 20 alignements sous-optimaux (figure 32-B, κ=1..20). On constate que la séquence peut se diviser en deux parties : les 63 premiers acides aminés, dont l'alignement est extrêmement variable, et les 24 derniers acides aminés, dont l'alignement est correct et ne varie pas. Ces 24 derniers acides aminés correspondent à la partie C-terminale de la séquence qui inclut les trois derniers éléments de structures secondaires, c'est-à-dire les brins β5 β6 et l'hélice α7. Cette coïncidence entre absence de variation et alignement local correct est remarquable et souligne que cette portion de séquence est probablement mieux conservée et plus facile à aligner que le reste de la séquence.

Parmi les alignements sous-optimaux générés on constate que certains alignements sont bien meilleurs que l'OSA. Les alignements κ=12, κ=15 et κ=16 sont les plus intéressants. Ils correspondent respectivement aux 12ème, 15ème et 16ème scores probabilistes les plus élevés dans le processus d'alignement de la séquence sur le HMM de la famille. Au sein de ces trois alignements, l'alignement correct des structures secondaires α_2, β_3 et α_4 est pratiquement rétabli. L'hélice α_2 est correctement alignée sur toute sa longueur. Le brin β_3 présente un décalage de 2 acides aminés par rapport à sa position correcte. Ce type de décalage est assez courant dans le positionnement des brins car ce sont des structures de période 2. Enfin, l'hélice α_4, qui dans l'OSA était alignée en face des positions de l'hélice α_2, est bien alignée sur toute sa longueur à l'exception de l'extrémité N-terminale. Le rétablissement de l'alignement correct de ces trois structures secondaires (à un décalage de 2 près pour le brin β_3) semble être corrélé. Ces trois évènements n'apparaissent en effet que simultanément.

Alignements sous-optimaux avec la méthode restreignant les appariements aux régions des structures secondaires. Nous avons effectué la même analyse que précédemment en utilisant la méthode de construction du modèle de Markov caché basée sur la conservation des structures secondaires (paragraphe 3.2.2). Là encore, 20 alignements sous-optimaux ont été générés grâce à la fonction HMMKALIGN (figure 32-C). Tout d'abord, on note que comme précédemment, il n'y a pas de variation dans l'alignement dans la région C-terminale, les structures secondaires β_5, β_6 et α_7 étant systématiquement correctement alignées. Dans l'alignement κ=7, seules deux structures secondaires sont mal positionnées : il s'agit des brins β_1 et β_3. Plus précisément, on constate que le brin β_1 est simplement décalé d'un acide

aminé, alors que dans l'OSA comme dans les 20 alignements sous-optimaux générés avec la méthode de construction du HMM traditionnelle, cette structure secondaire n'était jamais alignée sur le domaine thiorédoxine. Le brin β_3 est quant à lui décalé de 2 résidus. Dans l'alignement κ=12, l'erreur commise sur la structure β_3 est corrigée : une seule structure secondaire est donc mal positionnée, il s'agit de β_1. A l'inverse, dans l'alignement κ=19, le brin β_1 est correctement positionné et la seule structure secondaire décalée est β_3.

Conclusions. Au final, alors que le Q_{mod} de l'alignement de séquence optimal (OSA) est de 0.50, celui-ci atteint 0.79 lorsque l'on considère les 20 meilleurs alignements avec une construction du HMM traditionnelle, et 0.89 lorsque la méthode de construction du HMM est contrainte par la position des structures secondaires.

Ces résultats se reflètent au travers des modèles qu'on peut construire par homologie sur la base des différents alignements. La figure 32-D présente d'une part la structure native de la forme oxydée de la glutarédoxine du bactériophage T4, mais aussi les meilleurs modèles générés avec l'alignement de séquence optimal, l'alignement sous-optimal κ=16 lorsque le HMM est construit de manière traditionnelle et l'alignement κ=7 lorsque le HMM est construit en se basant sur la conservation des structures secondaires. On constate assez nettement que le dernier modèle est le meilleur.

3.4 Comparaison des améliorations obtenues avec HMMKALIGN et des améliorations obtenues en utilisant des méthodes d'alignements profil-profil.

La stratégie d'alignement profil-profil a été appliquée sur notre base test de 155 alignements très divergents. Nous avons utilisé HHPRED (Soding et al., 2005) pour aligner les 115 séquences sur le profil des autres membres de leur famille ; en incluant les informations de structures secondaires (prédites pour la séquence test, et réelles pour les séquences du profil de la famille). Les détails sont indiqués dans l'**annexe A**. Notons que pour 4 séquences, nous n'avons pas été en mesure d'obtenir un alignement profil-profil avec HHPRED : tous les résultats que nous présentons portent donc sur 111 séquences.

Dans un premier temps, le nombre d'améliorations obtenues en utilisant HHPRED plutôt que HMMer dans sa version traditionnelle a été calculé et comparé aux améliorations que l'on peut obtenir en utilisant HMMKALIGN. Les résultats, présentés dans la **table 9**, sont donnés en fonction du score Q_i pris en compte et du seuil ΔQ_i considéré.

Pour les valeurs de seuil ΔQ_i faibles, l'utilisation de HMMKALIGN est significativement plus efficace que celle de HHPRED quel que soit le score considéré (Q_{dev}, Q_{mod} et Q_{local}). Par exemple, pour un $\Delta Q_{local} > 0.01$, HMMKALIGN génère de meilleurs alignements que HMMer dans 90 cas sur 111 (81%), alors que HHPRED est plus efficace qu'HMMer dans 69 alignements sur 111 (62%). Lorsqu'on s'intéresse aux améliorations plus importantes, en fixant un seuil de $\Delta Qi > 0.10$, les résultats sont plus nuancés et dépendant des scores considérés. Pour le Q_{mod} et le Q_{dev}, les deux méthodes ont des résultats quasi-équivalents, autour de 25% d'améliorations. Le Q_{local} présente une tendance différente, puisque HHPRED donne de meilleurs résultats que HMMKALIGN (respectivement 24 alignements sur 111, soit 21% contre 16 sur 111, soit 14%).

Enfin, nous avons comparé la moyenne des Q_{mod}, Q_{dev} et Q_{local} obtenus sur les 111 alignements par les 3 approches :

(i) l'alignement de séquence optimal HMMer ;

(ii) l'exploration du voisinage de l'OSA par HMMKALIGN (20 alignements avec un HMM construit traditionnellement + 20 alignements avec un HMM dont les états d'appariements sont contraints par les régions des structures secondaires) ;

(iii) l'alignement de séquences profil-profil par HHPRED.

Les résultats sont présentés dans la **table 10**. On y constate que les résultats obtenus par HMMKALIGN et HHPRED sont très proches. Il n'est en effet pas possible de démontrer par un test de Student que ces distributions sont différentes, même en tolérant une erreur de 5%.

ΔQ_i	Q_{mod}	Q_{dev}	Q_{local}
Nombre d'alignements améliorés par HHPred			
0.00	66/111 (59%)	57/111 (51%)	70/111 (63%)
0.01	65/111 (59%)	56/111 (50%)	69/111 (62%)
0.02	57/111 (51%)	51/111 (46%)	59/111 (59%)
0.05	42/111 (38%)	39/111 (35%)	41/111 (41%)
0.10	27/111 (24%)	25/111 (23%)	24/111 (21%)
Nombre d'alignements améliorés par HmmKalign (K=20; avec l'une ou l'autre des deux méthodes de construction du HMM)			
0.00	91/111 (82%)	93/111 (84%)	92/111 (83%)
0.01	89/111 (80%)	92/111 (83%)	90/111 (81%)
0.02	63/111 (57%)	74/111 (67%)	60/111 (54%)
0.05	38/111 (34%)	48/111 (43%)	34/111 (31%)
0.10	25/111 (23%)	28/111 (25%)	16/111 (14%)

table 9 : Comparaison du Q_{mod}, Q_{dev} et Q_{local} de l'OSA et des alignements HHPred (en haut) et HMM-Kalign (en bas). Pour HMM-Kalign, 20 alignements ont été générés avec la méthode traditionnelle de construction du HMM, et 20 autres avec la méthode basée sur la conservation des structures secondaires. La première colonne indique le seuil *s* utilisé (seuil strict : $\Delta Q_i > s$). Dans la seconde colonne, on dénombre les alignements où ΔQ_i du Q_{mod} est supérieur ou égal au seuil *s*. Les résultats sont donnés en nombre d'occurrences (sur 111) et en pourcentage de l'ensemble de la base de test. Les colonnes suivantes indiquent les mêmes résultats respectivement pour le Q_{dev} et le Q_{local}.

	Qmod	Qdev	Qlocal
Moyenne et Ecart-type des 111 OSA			
moyenne	0.790	0.769	0.829
écart type	0.179	0.188	0.166
Moyenne et Ecart-type en utilisant HmmKalign (20 alignements sous-optimaux avec chaque méthode de construction du HMM)			
moyenne	0.847	0.838	0.880
écart type	0.148	0.156	0.136
Moyenne et Ecart-type des 111 alignements HMM-HMM			
moyenne	0.834	0.780	0.874
écart type	0.189	0.190	0.187

table 10 : Scores moyens et écarts types du Q_{mod}, Q_{dev} et Q_{local}. On compare (1) les 111 OSA classiquement générés par HMMer ; (2) les 111 meilleurs alignements existant dans le voisinage de l'OSA lorsqu'on explore les 20 alignements sous-optimaux générés sur le HMM classique + les 20 alignements sous-optimaux générés sur le HMM dont la construction est basée sur la conservation des éléments de structures secondaires ; (3) les 111 alignements générés par HHPred.

3.5 Discussion et perspectives de ce travail sur les alignements.

3.5.1 HmmKalign : une méthode de génération d'alignements alternatifs novatrice.

Le problème de la génération d'alignements alternatifs n'avait jusque récemment été abordé que dans le cadre des alignements séquence-séquence. Des travaux très récents ont proposé la ré-introduction de cette problématique dans le contexte des alignements profil-profil et séquence-structure (Chivian and Baker, 2006; Jaroszewski et al., 2002). Deux approches heuristiques ont été testées :

- Lukasz Jaroszewski, Weizhong Li et Adam Godzik ont ré-introduit au sein d'une méthode d'alignement profil-profil l'algorithme développé par Saqi et Sternberg en 1991 (*Iterative Elimination Method*). Pour une exploration encore plus importante de l'espace des séquences, cette méthode a été couplée à une heuristique paramétrique. Ainsi, pour chacune des 256 combinaisons de paramètres, 1000 alignements sont produits avec l'*Iterative Elimination Method* ; au final, 256000 alignements sont donc produits ! Les auteurs mettent en évidence que cette approche permet dans 48% des cas de générer au moins un alignement parmi les 256000 qui soit plus précis que l'alignement optimal classique.
- Dylan Chivian et David Baker ont introduit K*SYNC, un programme d'alignement séquence-structure au sein duquel la procédure classique de programmation dynamique est perturbée par la variation des paramètres en fonction des insertions/délétions, des prédictions de structures secondaires, etc… Pour produire encore davantage d'alignements, différentes matrices de substitutions sont utilisées. Ainsi pour aligner deux séquences du concours CASP, les auteurs produisent plus de 46000 alignements qu'ils réduisent à un ensemble non redondant comptant en moyenne 2500 alignements (Chivian and Baker, 2006). Cette opération est très coûteuse en temps : environ 10 minutes par couple de séquences sur un cluster de 54 processeurs AMD Athlon MP1600+.

Ces travaux n'ont cependant pas donné lieu à la mise à disposition de la communauté scientifique d'un programme permettant de réaliser ces alignements alternatifs.

La fonction HMMKALIGN implémentée au sein de HMMer est la première fonction permettant de générer un ensemble d'alignements alternatifs distribuée librement. Elle présente plusieurs avantages comparée aux études précédentes. Tout d'abord, l'ensemble des κ alignements générés est constitué des κ alignements dont les scores probabilistes sont les plus élevés. Ainsi, l'exploration est focalisée sur les alignements les plus probables. De plus, son exécution est particulièrement rapide. A titre d'exemple, il est possible de produire jusqu'à 800 alignements sous-optimaux, tous distincts et avec des scores néanmoins élevés, en moins d'une minute pour une séquence de 130 acides aminés de long.

3.5.2 Comparaison avec les autres méthodes de génération d'alignements alternatifs dans le cadre des alignements séquence-profil.

La publication décrivant le développement de HmmKalign est en cours de révision pour la revue Bioinformatics **(Article 2)**. Une remarque d'un des *referee* nous a conduit à comparer la fonction HmmKalign aux approches heuristiques citées dans la section précédente pour générer des alignements alternatifs. Pour cela, nous avons implémenté la méthode paramétrique et l'*Iterative Elimination Method* au sein de HMMer.

- L'approche paramétrique classique consiste à faire varier les paramètres *(u,v)* gérants les pénalités d'ouverture et d'extension des insertions et délétions pour produire des alignements alternatifs. Pour adapter la méthode paramétrique dans le cadre rigoureux des modèles de Markov cachés, nous avons implémenté une fonction qui modifie les probabilités de transitions pour favoriser (ou défavoriser) les transitions vers les états d'insertions et de délétions au détriment (ou à l'avantage) des transitions entre états d'appariement successifs.

- Dans le cadre des alignements par paires, l'*Iterative Elimination Method* consiste à modifier la matrice de similarité pour défavoriser toutes les cellules par lesquelles passe l'alignement optimal. Dans le cadre des modèles de Markov cachés, il nous a semblé que la solution la plus adéquate pour implémenter cette méthode était d'écrire une fonction qui modifie les probabilités d'émission des états d'appariement, de façon à ce que si dans l'alignement de séquence optimal l'acide aminé *x* a été émis

dans l'état $q_{i,}$ alors la probabilité d'émettre x dans l'état q_i diminue tandis que celle des autres acides aminés augmente.

Ces fonctions ont été implémentées au sein de HMMer et nous étudions actuellement les résultats de ces approches. Il nous faut tout d'abord choisir de façon adéquate les différents paramètres utilisés dans ces approches, c'est-à-dire l'amplitude des pénalités à appliquer aux probabilités de transition (Méthode Paramétrique) et d'émission (*Iterative Elimination Method*). Nous comparerons alors les résultats à ceux de la fonction HMMKALIGN sur les 115 alignements hautement divergents qui constituent notre base de test.

3.5.3 Le problème de la discrimination entre alignements corrects et incorrects.

Il serait intéressant de mettre au point une stratégie d'évaluation des alignements alternatifs obtenus qui soit capable de discriminer les alignements corrects (voir **figure 33**). Pour cela, la méthode la plus adéquate nous semble être de produire un modèle structural à partir de chaque alignement sous-optimal, et d'évaluer ce modèle à l'aide de fonction de scores classiques (PROSA, ANOLEA, MAXSUB, etc…). On peut en effet supposer que lorsque l'alignement structural fait partie des alignements sous-optimaux proposés par HMMKALIGN, les modèles produits à l'aide de cet alignement sont mieux évalués que les autres.

Ce procédé de sélection présente néanmoins des inconvénients. Lorsque l'alignement structural exact ne fait pas partie de l'ensemble des alignements sous-optimaux proposés, le score d'évaluation global ne permet pas de discriminer un alignement proche de l'alignement structural des autres alignements qui en sont très éloignés. Cette absence de gradient entre la qualité progressive de l'alignement et les scores de l'évaluation structurale nous a conduit à explorer des stratégies d'optimisation complémentaires. Parmi celles envisagées, une approche itérative s'inspirant du protocole qui suit est en cours d'étude (1) calcul des scores d'évaluation pour un ensemble de modèles le long de la séquence pour identifier les régions mal alignées (2) seules les régions *a priori* mal alignées sont soumises à une exploration « sous-optimale » (3) suite à la génération de nouveaux modèles, les améliorations observées permettent de bloquer de proche en proche les régions bien alignées. L'exploration ciblée d'une sous partie d'un alignement est déjà possible avec HmmKalign (cf article 2). Reste à élaborer l'algorithme qui permet de coupler l'évaluation structurale des modèles et la génération de contraintes sur l'alignement.

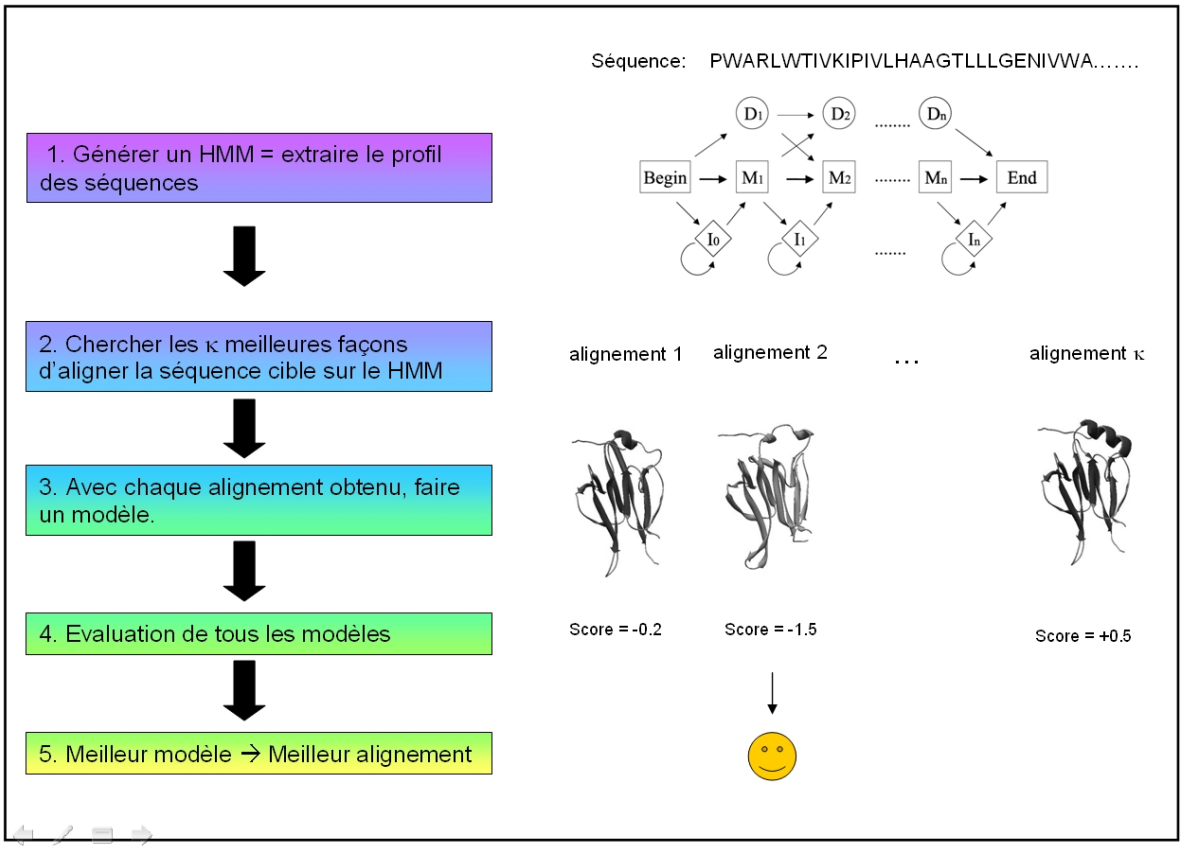


figure 33 : Procédure permettant de discriminer automatiquement les alignements corrects et incorrects. **(1)** Dans un premier temps, on paramétrise le HMM sur lequel on souhaite aligner la séquence cible. **(2)** A l'aide de la fonction HmmKalign, on produit κ alignements sous-optimaux. **(3)** Avec chacun de ces alignements, on produit plusieurs modèles par homologie. **(4)** Ces modèles sont évalués, et **(5)** on identifie le meilleur modèle, que l'on suppose provenir du meilleur alignement.

3.5.4 Adaptation de HmmKalign aux alignements HMM-HMM ?.

La fonction HMMKALIGN utilise l'algorithme de Viterbi généralisé permettant de trouver la meilleure façon d'aligner une séquence sur un HMM donné, et s'applique donc naturellement dans le cadre des alignements séquence-profil. Nous envisageons pour des travaux futurs d'implémenter ce même algorithme dans le cadre de certaines méthodes de comparaison profil-profil dont le formalisme sous-jacent est basé sur les modèles de Markov caché, comme HHPRED par exemple. Il nous faudra cependant attendre que les codes sources de ces programmes soient librement distribués.

Chapitre 4 : Détection des sites de liaison des PRMs sur la séquence de leurs partenaires. Application aux interactions FHA – partenaires des voies de surveillance des dommages de l'ADN.

Ce chapitre s'intéresse à la détection des sites reconnus par les PRMs sur la séquence de leurs partenaires. Une stratégie prédictive a été mise au point pour détecter les thréonines spécifiquement reconnues par les domaines FHA et a été appliquée au réseau d'interactions protéine-protéine mis en place suite à une cassure double brin de l'ADN chez Saccharomyces cerevisiae. Nous avons collaboré avec les équipes de Françoise Ochsenbein et Marie-Claude Marsolier-Kergoat qui s'intéressent toutes deux à la protéine kinase Rad53.

4.1 La protéine Rad53, kinase essentielle des voies de surveillance des dommages de l'ADN.

4.1.1 Aspects structuraux de Rad53.

La protéine Rad53 a été identifiée en 1991 chez *Saccharomyces cerevisiae* (Stern et al., 1991) grâce aux travaux de l'équipe de Cynthia Zerillo (*Yale University School of Medecine*, New Haven, USA). Initialement appelée Spk1, cette protéine de 821 résidus a été décrite comme une kinase capable de phosphoryler les sérines, thréonines et tyrosines. Les premiers travaux associant Rad53 au contrôle du cycle cellulaire chez la levure sont parus en 1994 (Allen et al., 1994). Au cours des années suivantes, des protéines orthologues de Rad53 ont été mises en évidence au sein d'autres organismes. L'orthologue humain de Rad53 a été identifié en 1998 par l'équipe de Stephen Elledge (*Howard Hughes Medical Institute*, Houston, USA). Cette kinase, assez divergente en terme de séquence avec Rad53, a été nommée Chk2 (Chaturvedi et al., 1999; Matsuoka et al., 1998).

Les protéines de la « famille Chk2 » ont une composition en domaines bien conservée malgré une identité de séquence faible (**figure** 34) : toutes possèdent une région N-terminale riche en répétitions sérines-glutamines/thréonines-glutamines (SQ/TQ), suivie d'un domaine FHA (identité de séquence moyenne 20%) et d'un domaine kinase (identité de séquence moyenne 40%). La protéine Rad53 se singularise des autres membres de la famille par la présence d'une large extension C-terminale comprenant un second domaine FHA. Seuls Rad53 et ses orthologues chez les autres levures *Saccharomyces* possèdent ce second domaine FHA.

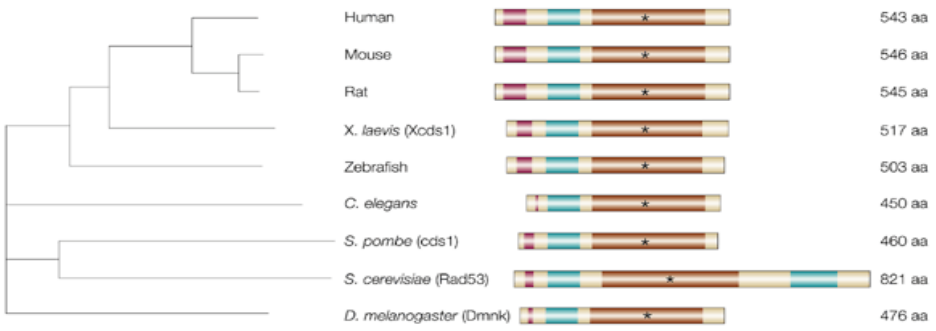


figure 34 : Les protéines de la «famille Chk2 » représentées sur un arbre phylogénétique. Le nom des différentes protéines est représenté entre parenthèses. Les protéines sont composées d'une région riche en SQ-TQ (en violet), d'un ou plusieurs domaines FHA (en bleu) et d'un domaine kinase (en brun), dont la boucle d'activation est symbolisée par une étoile. Cette figure est extraite de (Bartek et al., 2001).

Les délimitations exactes des différents domaines de Rad53 sont présentées sur la **figure** 35. Les structures des deux domaines FHA sont connues, tandis que la structure du domaine kinase central n'a pas encore été résolue expérimentalement. Néanmoins, la structure du domaine kinase de Chk2 est disponible depuis 2006 (Oliver et al., 2006).

figure 35 : Délimitations des domaines de la protéine Rad53 : les répétitions SQ-TQ (en violet), les deux domaines FHA (en vert) et le domaine kinase (en rouge). Les numéros correspondant aux premiers et derniers résidus de chaque domaine sont notés sur la ligne du dessous. Ces délimitations sont extraites de (Bartek et al., 2001).

La **figure 36**-A représente la structure du domaine FHA N-terminal de Rad53, également appelé domaine FHA1, résolue en 2000 par diffraction des rayons X au sein de l'équipe de Stephen Jackson (*Institute of Cancer and Developmental Biology*, Cambridge, UK). Dans la même publication (Durocher et al., 2000), il a été établi que le domaine FHA1 de Rad53 reconnaît des peptides et des fragments protéiques contenant une thréonine phosphorylée (pT) préférentiellement suivie d'un aspartate en position pT+3 (Kd = 330nM). Cette propriété a été déduite du criblage d'une bibliothèque de courts peptides phosphorylés. Les substitutions en alanine de deux positions très conservées du FHA, l'arginine 70 et la sérine 85, abolissent totalement l'interaction entre le domaine FHA1 de Rad53 et les peptides ou fragments protéiques phosphorylés (Durocher et al., 1999). La structure du domaine FHA1 de Rad53 en interaction avec un peptide phosphorylé permet d'expliquer cette perte d'interaction : la chaîne latérale de l'arginine 70 et celle de la sérine 85 sont en contact direct avec le groupement phosphate de la thréonine phosphorylée. La **figure 36**-B illustre le réseau des contacts entre le domaine FHA1 de Rad53 et un peptide phosphorylé respectant le motif de plus haute affinité, pTxxD.

L'équipe de Ming-Daw Tsai (*Ohio State University*, Columbus, USA) a étudié la structure du domaine FHA C-Terminal de Rad53, souvent nommé domaine FHA2. En 1999, ils ont résolu la structure de ce domaine, non complexé, par résonance magnétique nucléaire (Liao et al., 1999). Dans les deux années suivantes, il a été établi que le domaine FHA2 de Rad53 se lie *in vitro* à des peptides phosphorylés sur des thréonines (Durocher et al., 2000), mais également, chose plus étonnante, sur des tyrosines (Liao et al., 1999; Wang et al., 2000). Les structures correspondant à ces complexes sont présentées sur la **figure 37**.

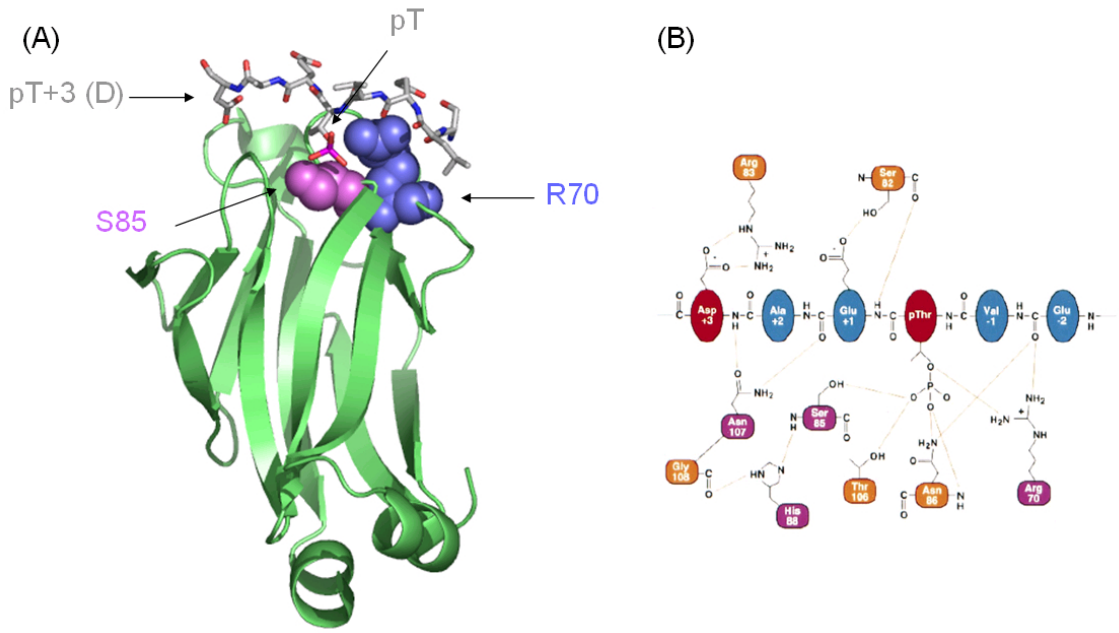


figure 36 : (A) Structure résolue par diffraction des rayons X du domaine FHA1 de Rad53 en interaction avec un peptide phosphorylé respectant le motif pTxxD (code PDB 1G6G). Le domaine FHA1 de Rad53 est représenté par un ruban vert. Les deux résidus dont la substitution en alanine abroge l'interaction sont représentés en sphères bleues pour l'arginine 70 et roses pour la sérine 85. Le peptide phosphorylé est représenté en gris. (B) Représentation schématique des contacts lors de l'interaction (ovale = peptide ; rectangle = domaine FHA). Figure extraite de (Durocher et al., 2000).

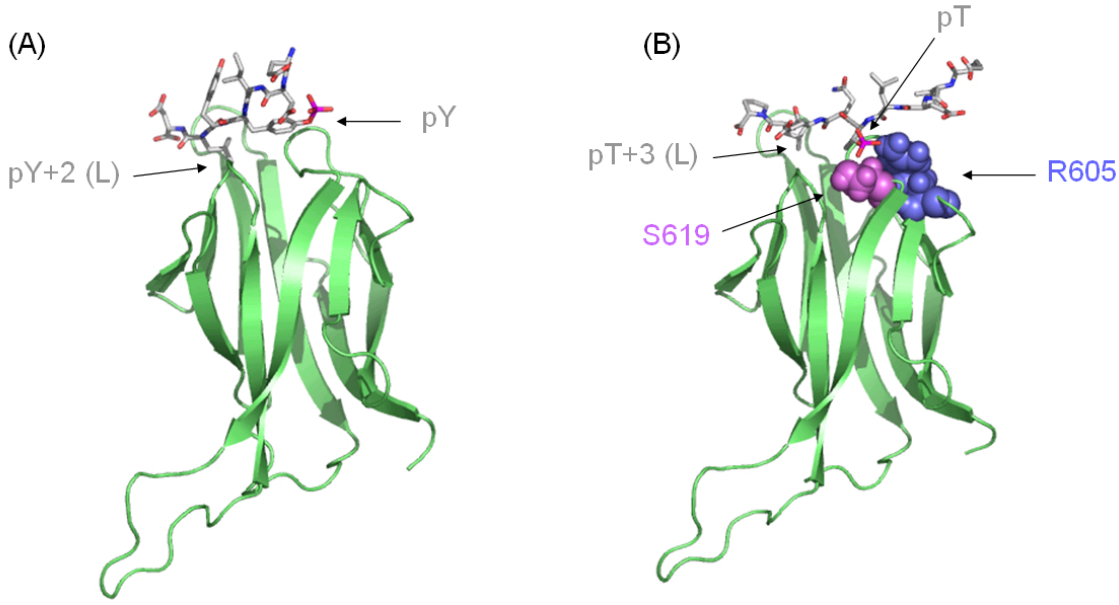


figure 37 : (A) Structure tridimensionnelle (RMN, structure raffinée) du domaine FHA2 de Rad53 en interaction avec un peptide phosphorylé respectant le motif pYxL (code PDB 1J4K). Le domaine FHA2 de Rad53 est représenté par un ruban vert. Le peptide phosphorylé est représenté en gris. (B) Structure tridimensionnelle (RMN, structure raffinée) du domaine FHA2 de Rad53 en interaction avec un peptide phosphorylé respectant le motif pTxxL (code PDB 1J4L). Le domaine FHA2 de Rad53 est représenté par un ruban vert. Les deux résidus dont la substitution en alanine abroge l'interaction sont représentés en sphères bleues pour l'arginine 605 et roses pour la sérine 619. Le peptide phosphorylé est représenté en gris.

La sélectivité du domaine FHA2 de Rad53 pour ses substrats est moins importante que celle du domaine FHA1 de Rad53 puisqu'il peut reconnaître des motifs associant une thréonine phosphorylée (pT) et une isoleucine ou une leucine en position pT+3 (Byeon et al., 2001; Durocher et al., 2000), mais aussi des motifs associant une tyrosine phosphorylée (pY) à une leucine en position pY+2 (Wang et al., 2000). Cependant, l'affinité de l'interaction avec les motifs contenant des pY (Kd = 4 mM) est nettement inférieure à celle de l'interaction avec les motifs contenant des pT (Kd = 10 μM). De plus, il existe peu de tyrosines kinases chez *S. cerevisiae*. La conjugaison de ces deux facteurs suggère que l'interaction avec le motif pY a peu de vraisemblance physiologique.

Pour assurer la liaison du domaine FHA2 de Rad53 aux peptides contenant des thréonines phosphorylées, l'arginine 605 et la sérine 619 jouent le même rôle que l'arginine 70 et la sérine 85 dans le domaine FHA1. La substitution de l'un ou l'autre de ces résidus en alanine abroge totalement l'interaction du domaine FHA2 avec ces peptides (Durocher et al., 1999).

4.1.2 Rôle de la protéine Rad53 dans la signalisation des dommages de l'ADN.

Les cellules sont fréquemment soumises à des stress endogènes ou exogènes endommageant l'ADN. Une réparation fidèle de ces lésions est nécessaire pour le maintien de l'intégrité du matériel génétique et pour la survie cellulaire. Les défauts de réparation peuvent en effet constituer des éléments d'initiation de la cancérogenèse (Rouse and Jackson, 2002).

Parmi toutes les lésions possibles de l'ADN, les cassures double brin sont probablement les plus dangereuses. Suite à une cassure double brin, une cascade d'interactions protéine-protéine se met en place au sein de la cellule. La lésion doit tout d'abord être identifiée précisément, puis le signal doit être transmis et amplifié afin de déclencher la réponse adéquate, qui peut être l'arrêt temporaire de la progression du cycle cellulaire, l'expression des gènes participant à la réparation de la lésion, ou dans le cas des eucaryotes supérieurs l'induction de la mort cellulaire programmée (Khanna and Jackson, 2001).

Au sein de la cascade d'interactions protéine-protéine mise en place suite à une cassure double brin, la kinase Rad53 joue un rôle central. Après que la lésion a été détectée, Rad53 est activée par phosphorylation par l'intermédiaire des protéines Rad9 et Mec1 (**figure 38**). L'activation de Rad53 déclenche alors son autophosphorylation et la protéine devient

hyperphosphorylée (Sweeney et al., 2005). La déphosphorylation de Rad53 est nécessaire pour que la progression du cycle cellulaire reprenne (Pellicoli et al., 2001 ; Vaze et al.,2002), lorsque les lésions ont été réparées (processus de rétablissement), ou lorsque malgré la présence de dommages, certaines cellules reprennent la progression du cycle (processus d'adaptation).

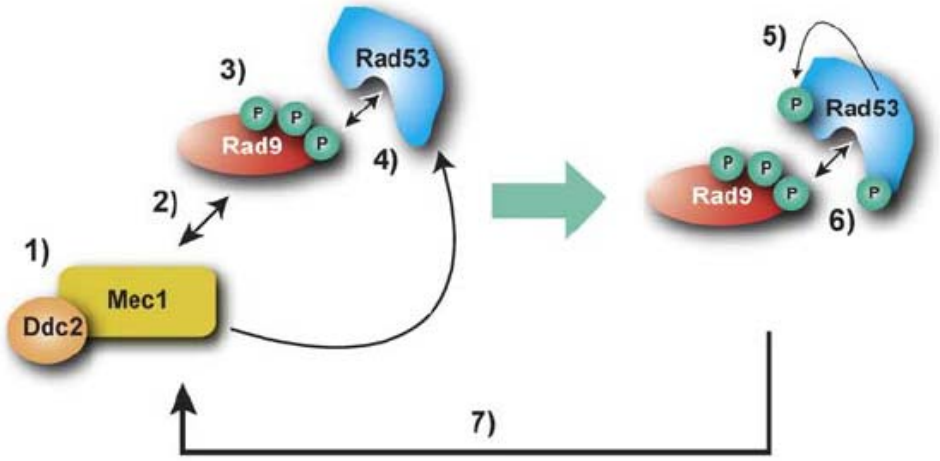


figure 38 : Modèle en 6 étapes d'activation de Rad53 suite à une cassure double brin de l'ADN, proposé par l'équipe de Daniel Durocher (Sweeney et al., 2005). Suite au dommage : (1) activation de Mec1 et Ddc2 ; (2) phosphorylation de Rad9 sur de multiples sites par Mec1 ; (3) recrutement de Rad53 par Rad9 ; (4) phosphorylation de Rad53 par Mec1 par l'intermédiaire de Rad9. Par la suite, (5) Rad53 s'autophosphoryle avant de (6) se dissocier de Rad9. L'étape (4) est une hypothèse des auteurs, les mécanismes exacts d'activation de Rad53 par Mec1 et/ou Rad9 ne sont pas complètement élucidés à l'heure actuelle (février 2007).

4.1.3 Partenaires connus de Rad53.

De par son rôle central dans les voies de signalisation des dommages de l'ADN, Rad53 interagit avec un grand nombre de partenaires. A titre d'exemple, la base de données BIND (Gilbert, 2005) référence plus d'une dizaine d'interactions impliquant Rad53 (**table 11**).

Rad53 fait donc partie des protéines dont le degré de connexité au sein des réseaux d'interactions protéine-protéine est très élevé : c'est un *hub*. Parvenir à étudier indépendamment toutes les interactions de Rad53 afin de comprendre précisément le rôle de chacune d'entre elles est particulièrement intéressant. La majorité des interactions entre Rad53 et ses partenaires est médiée par l'un des deux domaines FHA de Rad53. C'est notamment le cas des interactions avec Rad9, Ptc2, Dbf4 et Asf1.

Partenaire	Première(s) publication(s) associée(s)
Mec 1	(Desany et al., 1998; Lee et al., 2003; Sweeney et al., 2005)
Rad53	(Lee et al., 2003)
Asf1	(Emili et al., 2001)
Rad9	(Durocher et al., 1999; Emili, 1998; Liao et al., 2000; Vialard et al., 1998)
Dbf4	(Dohrmann et al., 1999)
Dun1	(Lee et al., 2003)
Hho1	(Lee et al., 2003)
Ptc2	(Leroy et al., 2003)
Swi6	(Sidorova and Breeden, 2003)
Sgs1	(Bjergbaek et al., 2005)
Rnr4	(Huang and Elledge, 1997)
Complexe Cdc7/Dbf4	(Kihara et al., 2000)
Stn1	(Ito et al., 2001)

table 11 : Partenaires connus de Rad53 d'après la base de données BIND (le code associé à Rad53 dans BIND est 6325104). La première colonne indique le nom du partenaire. Les publications associées sont indiquées dans la seconde colonne.

4.2 Détection efficace des sites reconnus par le domaine FHA1 de Rad53.

4.2.1 Approche croisée : conservation, phosphorylabilité, respect du motif spécifique.

Notre groupe collabore avec deux chercheurs du CEA intéressés par la protéine Rad53. Dans notre équipe, Françoise Ochsenbein (iBiTec-S, Laboratoire de Biologie Structurale et Radiobiologie) étudie le couplage entre l'assemblage du nucléosome et la réponse aux stress génotoxiques. L'équipe de Marie-Claude Marsolier-Kergoat (iBiTec-S, Laboratoire du Contrôle du Cycle Cellulaire) étudie quant à elle les voies de signalisation associées aux *checkpoints* du cycle cellulaire.

Puisque Rad53 est une protéine interagissant avec un grand nombre de partenaires, il n'est pas possible d'étudier finement sa fonction sans chercher à comprendre le rôle précis de

chaque interaction indépendamment les unes des autres. C'est la raison pour laquelle nous proposons, dans le cas d'une interaction entre un domaine FHA de Rad53 et un partenaire λ, de rechercher au sein de la séquence du partenaire λ la thréonine spécifiquement reconnue par le domaine FHA de Rad53 pour produire des mutants dans lesquels cette thréonine reconnue est substituée en alanine. De cette façon, il est possible d'étudier le rôle d'une interaction précise de Rad53 sans pour autant perturber le reste des interactions mettant en jeu Rad53 (**figure 39**).

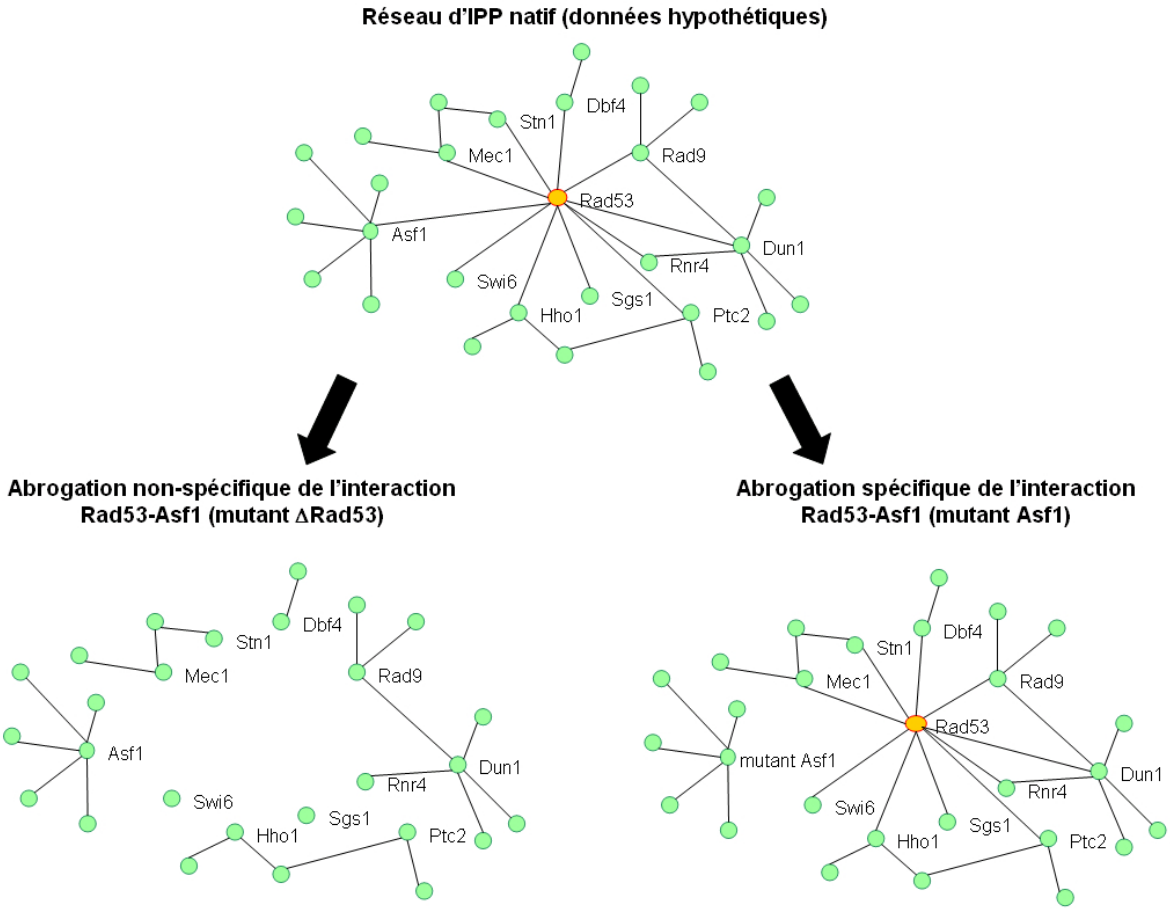


figure 39 : Réseau d'interactions protéine-protéine au voisinage de Rad53. Afin d'étudier finement le rôle de l'interaction entre Rad53 et Asf1, il est moins destructeur pour le réseau de produire un mutant de Asf1 ne se liant plus a Rad53 (a droite) que de produire un mutant ΔRad53 (à gauche). *Le réseau présenté est un réseau théorique : en raison de la complexité du réseau, les interactions mettant en jeu les partenaires de Rad53 ont été imaginées.*

Nous avons donc cherché à mettre en place une stratégie bioinformatique permettant de prédire la position des sites d'interaction des domaines FHA sur la séquence de leurs partenaires. Pour cela, plusieurs approches comprenant la conservation de la séquence du partenaire, la phosphorylabilité des thréonines, et le respect des motifs les plus affins pour les domaines FHA ont été explorées.

(i) Phosphorylabilité des thréonines. Etant donné que les domaines FHA reconnaissent des résidus phospho-thréonine, l'un des critères les plus naturels est de tester le caractère phosphorylable des thréonines de la séquence du partenaire.

Le réseau de neurones NETPHOS2.0 (Blom et al., 1999) a été développé dans l'équipe de Søren Brunak (*The Technical University of Denmark*, Lyngby, Danemark) afin d'identifier au sein d'une séquence protéique les sérines, thréonines et tyrosines susceptibles d'être phosphorylées *in vivo*. L'apprentissage du réseau de neurones a été effectué sur une banque de courts fragments protéiques contenant des sites de phosphorylation vérifiés expérimentalement (Kreegipuu et al., 1999) : 210 fragments contenant des tyrosines phosphorylées, 584 fragments contenant des sérines phosphorylées et 108 fragments contenant des thréonines phosphorylées. Les résultats publiés validant l'approche de NETPHOS2.0 montrent que 65% des sites contenant des thréonines phosphorylées sont prédits comme tels (sensibilité = 0.65), 52% des thréonines prédites comme phosphorylées le sont effectivement (précision = 0.52) ; et 83% des thréonines prédites comme non phosphorylées sont effectivement annotées comme telles (spécificité = 0.83).

(ii) Respect du motif le plus affin. Les motifs pour lesquels les domaines FHA de Rad53 sont les plus affins sont connus : le domaine FHA1 de Rad53 reconnaît préférentiellement les motifs pTxxD, tandis que le domaine FHA2 de Rad53 reconnaît les motifs pTxxI avec une sélectivité moins forte. Pour chaque thréonine de la séquence du partenaire il est donc rapide de tester si l'acide aminé en position +3 relativement à cette thréonine respecte le motif pTxxD ou pTxxI selon que l'interaction est médiée respectivement par le domaine FHA1 ou FHA2 de Rad53.

Le choix de ce critère soulève cependant une interrogation. En effet, des études de 2004 réalisées dans le groupe de Ming-Daw Tsai ont montré que dans le cas de l'interaction entre

le domaine FHA1 de Rad53 et Mdt1, la thréonine reconnue par le domaine FHA1 de Rad53 ne respecte pas la règle du motif le plus affin : pTxxD. A l'inverse, seule une publication fait état d'une interaction entre un domaine FHA et un fragment protéique respectant son motif de plus haute affinité (interaction FHA2 de Rad53 - Rad9, (Byeon et al., 2001; Durocher et al., 1999). Une discussion très intéressante amorcée par Ming-Daw Tsai remet en cause le modèle selon lequel les domaines FHA lieraient des motifs linéaires courts en reconnaissant sélectivement le résidu pT+3 (Mahajan et al., 2005). Nous avons inclus le critère du respect du motif court de plus haute affinité dans notre protocole d'analyse pour évaluer sa pertinence dans le contexte *in vivo*.

(iii) Conservation du site dans les espèces proches. Enfin, nous avons envisagé l'hypothèse d'une conservation du site reconnu par le domaine FHA dans les autres levures *Saccharomyces*, en nous concentrant sur la conservation des thréonines et des acides aminés en position +3 relativement aux thréonines. Pour estimer la conservation de ces résidus, nous utilisons la banque de données SGD (http://www.yeastgenome.org/; (Christie et al., 2004; Weng et al., 2003) qui référence les ORFs de *S. cerevisiae*, *S.paradoxus*, *S.mikatae, S.bayanus, S. castellii, S. kluyveri* et *S. kudriavzevii* (Cliften et al., 2003; Kellis et al., 2003). Ainsi, pour l'étude d'un partenaire λ de domaine FHA, on peut récupérer les séquences homologues sur le site de SGD, réaliser un alignement multiple et analyser la conservation de la séquence.

Le protocole d'analyse complet est présenté **figure 40**. Chaque séquence de partenaire d'un domaine FHA est analysée selon les trois critères cités précédemment. Pour chaque site d'interaction potentiel, nous obtenons une information booléenne de respect ou non du motif spécifique, ainsi qu'un taux de conservation variant entre 0 et 1 et une probabilité pour les thréonines d'être phosphorylées. Pour ces deux dernières informations, il faut utiliser une valeur seuil qui détermine si le site d'interaction potentiel respecte la contrainte.

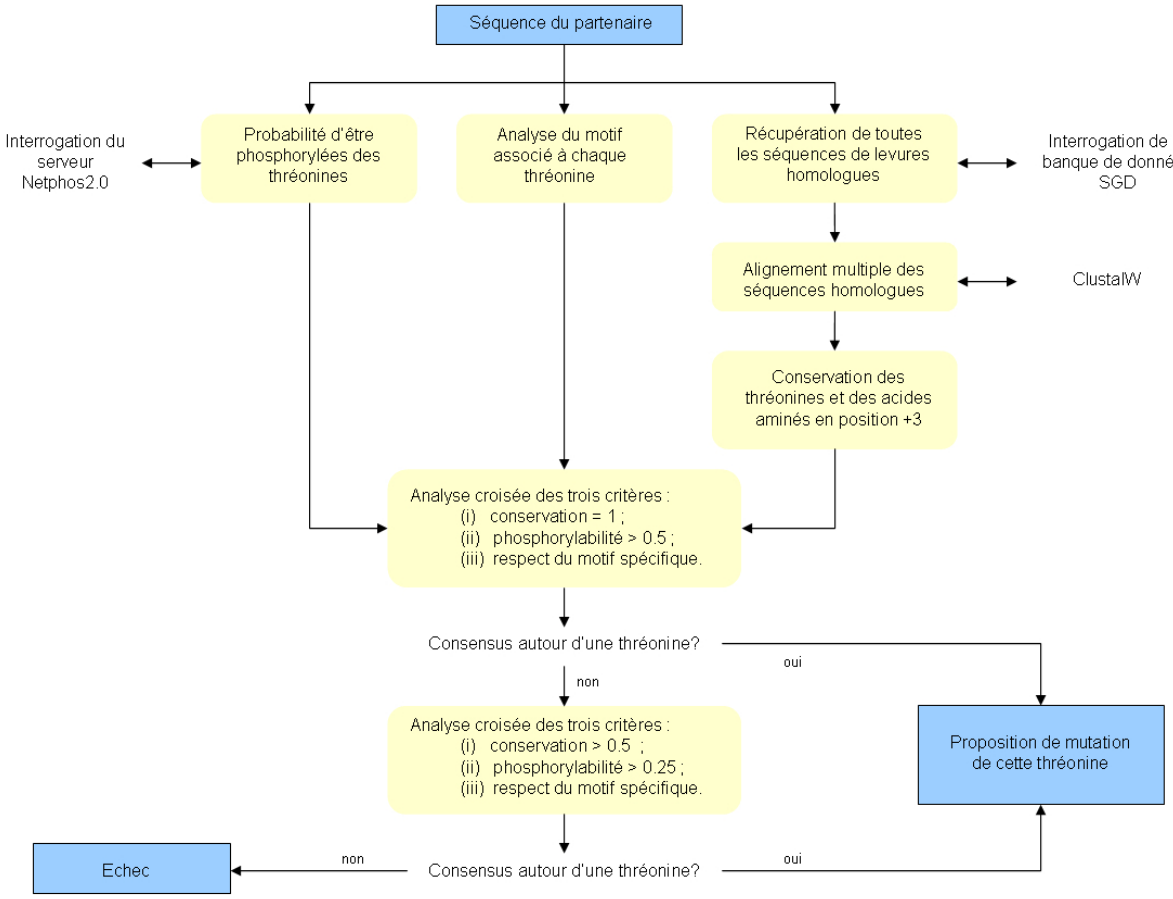


figure 40 : Identification au sein d'une séquence des thréonines pouvant être des sites d'interaction d'un domaine FHA. Dans un premier temps, la séquence est analysée par trois méthodes : la phosphorylabilité des thréonines est calculée par Netphos2.0 (le seuil par défaut du programme est fixé à 0,5) ; le respect du motif spécifique ; la conservation des thréonines et des acides aminés en position +3 relativement à ces thréonines. Si cette première analyse permet d'identifier une thréonine, on s'arrête là. Sinon, on effectue la même analyse en étant moins stringeant quant aux critères de phosphorylabilité et de conservation (chaque seuil est divisé par deux).

Plusieurs éléments nous ont conduit à assouplir le protocole mentionné ci-dessus : (i) une conservation parfaite entraîne une valeur seuil de 1, mais rien ne nous assure que le site d'interaction soit conservé ou ne puisse pas être décalé de quelques acides aminés dans deux espèces proches ; (ii) dans le programme NETPHOS-2.0, la valeur seuil par défaut ne permet de prédire que 65% des thréonines phosphorylées. Nous avons donc choisi une stratégie en deux étapes. La première consiste à chercher un consensus entre les différentes sources d'information en utilisant les seuils les plus stringeants : 1 pour la conservation, et 0.5 pour la phosphorylabilité (seuil par défaut de NETPHOS-2.0). Si aucun consensus ne se dégage, les valeurs seuils sont abaissées afin de chercher un site d'interaction consensuel.

Nous détaillons maintenant les interactions de Rad53 étudiées par Françoise Ochsenbein et Marie-Claude Marsolier-Kergoat sur lesquelles cette analyse croisée a été appliquée afin d'identifier la thréonine reconnue par Rad53. Pour les deux interactions Rad53-Ptc2 et Rad53-Asf1, la stratégie bioinformatique présentée a été appliquée avec succès et nous présenterons brièvement les résultats expérimentaux obtenus par nos collaborateurs.

4.2.2 Etude de l'interaction FHA1 de Rad53 – Ptc2.

L'équipe de Marie-Claude Marsolier-Kergoat a montré en 2003 (Leroy et al., 2003) que Rad53 et la phosphatase Ptc2 interagissent *in vitro* et *in vivo*. Cette interaction est médiée par le domaine FHA1 de Rad53 et est phospho-dépendante : les mutations R70A et S85A du domaine FHA1 de Rad53 entraînent toutes deux la perte de l'interaction. Dans cette publication, les auteurs proposent que la fonction de Ptc2 soit de déphosphoryler Rad53 de façon à permettre l'adaptation et le rétablissement des cellules. Ptc2 étant impliquée dans de nombreuses interactions, l'identification de la thréonine de Ptc2 spécifiquement reconnue par le domaine FHA1 de Rad53 était cruciale pour permettre de concevoir des mutants de Ptc2 incapables de se lier au domaine FHA1 de Rad53 et ainsi d'étudier la fonction de l'interaction Rad53-Ptc2 plus précisément.

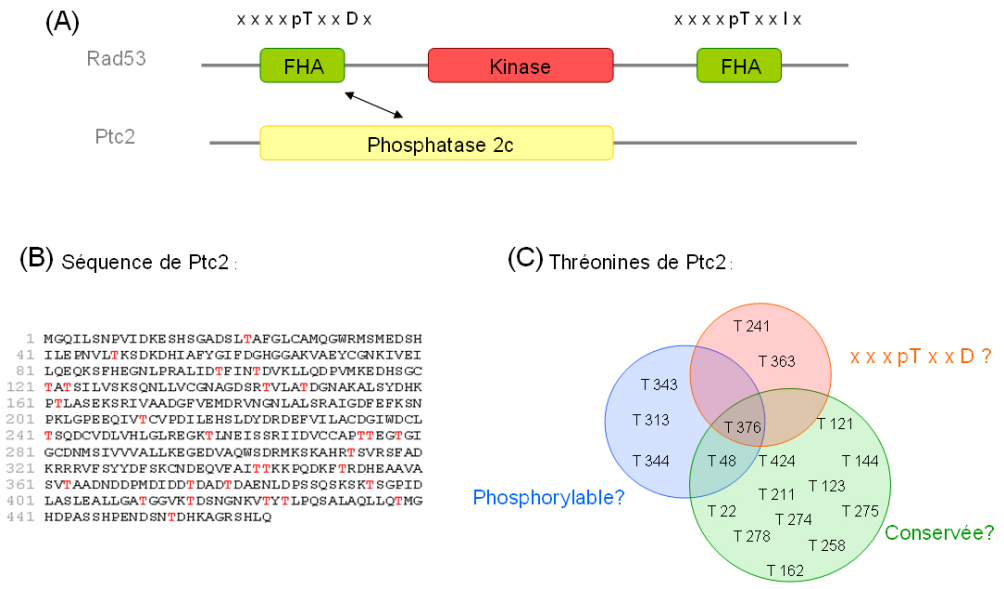


figure 41 : (A) Etude de l'interaction entre le domaine FHA1 de Rad53 et Ptc2. (B) Séquence complète de la protéine Ptc2. Les thréonines sont indiquées en rouge. (C) Analyse des différentes thréonines de la séquence de Ptc2 par les 3 critères de respect du motif de plus forte affinité du domaine FHA1 de Rad53, pTxxD (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 relativement à la thréonine (ensemble vert). Seule la thréonine T376 respecte ces trois critères simultanément (intersection des trois ensembles).

La séquence de PTC2 comprend 29 thréonines, en rouge dans la **figure 41**-B. Raphaël Guérois a construit un modèle par homologie du domaine phosphatase de la protéine PTC2, puis il a étudié l'accessibilité des quatre thréonines respectant le motif pTxxD sur le modèle. Il en a conclu que la thréonine T376 était une candidate sérieuse. En appliquant la stratégie conservation/phosphorylation/motif spécifique explicitée précédemment, j'ai également constaté que cette thréonine, qui fait partie de la boucle C-terminale du domaine phosphatase, était la meilleure candidate à la reconnaissance par le domaine FHA1 de Rad53 (**figure 41**-C). Effectivement, la thréonine 376 est la seule à respecter simultanément les trois critères imposés : sa séquence comporte un aspartate en position pT+3, son score de phosphorylabilité est élevé (score = 0.642), et la conservation de cette thréonine comme celle du motif qui la suit est parfaite (conservation = 1). L'analyse bioinformatique de la séquence entourant la thréonine 376, particulièrement acide, suggère que la Caséine Kinase 2 (CK2) pourrait être à l'origine de la phosphorylation de cette thréonine.

L'équipe de Marie-Claude Marsolier-Kergoat a montré que la substitution T376A entraîne *in vitro* une perte de l'interaction entre le domaine FHA1 de Rad53 et Ptc2. En effet, alors que le domaine FHA1 de Rad53 et Ptc2 interagissent par double hybride, cette interaction n'est pas retrouvée lorsque Ptc2 est muté en Ptc2_T376A (**figure 42**). Pour étayer ces résultats, Françoise Ochsenbein a réalisé une étude par RMN basée sur l'observation des variations de déplacements chimiques à partir du domaine FHA1 de Rad53 en absence et en présence d'un peptide ^{372}DDID(pT)DADTDAE^{383} dérivé de Ptc2. Les résultats obtenus sont cohérents avec les résultats de double hybride puisqu'ils indiquent que le domaine FHA1 de Rad53 interagit directement avec ce peptide phosphorylé issu de Ptc2 et incluant la thréonine T376 *in vitro*.

L'étude du phénotype associé aux mutants Ptc2_T376A révèle des défauts d'adaptation et de rétablissement suite à des cassures double brins de l'ADN. De plus, il a été mis en évidence que les sous unités régulatrices Ckb1 et Ckb2 de la caséine kinase 2 (CK2) sont nécessaires à l'interaction FHA1 de Rad53-Ptc2 *in vivo* ; que Ckb1 se lie à Ptc2 *in vitro* ; et que les mutants *ΔCkb1* et *ΔCkb2* présentent des défauts d'adaptation et de rétablissement. L'ensemble de ces résultats (Guillemain et al., 2007) suggèrent fortement que la kinase CK2 est responsable *in vivo* de la phosphorylation de Ptc2 sur la thréonine T376.

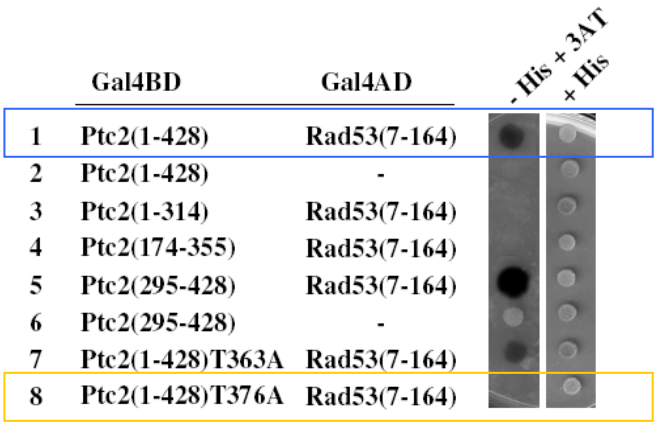


figure 42 : Résultats de double hybride. Le domaine FHA1 de Rad53 interagit avec Ptc2 et la thréonine T376 de Ptc2 est nécessaire à cette interaction. En bleu, l'interaction entre Ptc2 *wild-type* et le domaine FHA1 de Rad53. L'interaction est perdue avec le mutant T376A de Ptc2 (en jaune) (Guillemain et al., 2007). La colonne de droite permet de contrôler que les colonies sont viables.

4.2.3 Etude de l'interaction FHA1 de Rad53 – Asf1.

Asf1, pour *Anti-Silencing Function 1*, a été identifiée en 1997 chez *S. cerevisiae* et depuis de nombreuses études ont montré que les rôles d'Asf1 sont multiples. C'est une protéine chaperon d'histones impliquée dans l'assemblage du nucléosome (Adkins and Tyler, 2004; Mello et al., 2002; Tagami et al., 2004; Tyler et al., 1999). Elle intervient également dans la répression de la transcription de certains gènes (Le et al., 1997; Sharp et al., 2001; Tyler et al., 1999), dans le maintien de la stabilité du génome au cours de la réplication de l'ADN (Myung et al., 2003; Prado et al., 2004; Ramey et al., 2004), ou encore dans la réponse à différents stress génotoxiques (Emili et al., 2001; Hu et al., 2001; Recht et al., 2006; Verreault et al., 1996). Dans notre équipe, Françoise Ochsenbein travaille à la caractérisation structurale des interactions mettant en jeu Asf1. Ses travaux, auxquels j'ai collaboré, ont notamment permis de mettre en évidence les bases structurales de l'interaction entre Asf1 et le complexe d'histones H3/H4 (**Article 4**). La structure du complexe a ensuite été établi par Morgane Agez, doctorante de notre équipe (Agez et al., 2007).

Chez la levure *S. cerevisiae*, Rad53 interagit avec Asf1 (Emili et al., 2001). L'interaction entre ces deux protéines a été particulièrement étudiée pendant la thèse d'Aurélie Lautrette encadrée par Françoise Ochsenbein (Lautrette, 2006). Ce travail a permis de mieux caractériser le complexe formé par Rad53 et Asf1 chez *S. cerevisiae*, qui est médié par une

interaction entre le domaine FHA1 de Rad53 et l'extrémité C-terminale de Asf1. Les délimitations exactes du fragment interagissant sont indiquées sur la **figure 43**-A.

La partie C-terminale d'Asf1 contient 5 thréonines : T215, T220, T265, T270 et T278. Nous avons réalisé un alignement des séquences d'Asf1 chez les levures proches de *S. cerevisiae*. Ceci nous a permis de constater que l'extrémité C-terminale de Asf1 était plus divergente que le reste de la séquence. Parmi les quatre thréonines de ce fragment, T270 est bien conservée : elle est présente dans 8 séquences sur 9. Les résidus suivant T270, et plus particulièrement le résidu en position +3, sont également bien conservés. De plus, T270 est prédite comme phosphorylable (avec un score supérieur à 0.99) par le programme NETPHOS2.0 et respecte le motif pTxxD qui est particulièrement affin pour le domaine FHA1 de Rad53.

La combinaison de ces différents éléments fait de T270 un site d'interaction potentiel pour le domaine FHA1 (**figure 43**-C). Dans le but d'apporter des arguments expérimentaux, Aurélie Lautrette a conçus deux double mutants d'Asf1 : le double mutant Asf1 T215A+T220A, et le double mutant Asf1 T265A+T270A. Elle a montré par GST-pulldown (**figure 44**) que l'interaction entre le domaine FHA1 de Rad53 et le double mutant Asf1 T215A+T220A est détectable, tandis qu'elle ne l'est pas dans le cas du double mutant Asf1 T265A+T270A.

En résumé, les résultats suggèrent que le domaine FHA1 de Rad53 interagit avec l'extrémité C-terminale de Asf1 en reconnaissant la thréonine T270. Une étude de Marcus Smolka (*Ludwig Institute for Cancer Research, San Diego, USA*) réalisée au courant de l'année 2005 dans laquelle les sites de phosphorylation de Rad53 et ses partenaires ont été analysés par spectrométrie de masse, mentionne la présence d'un peptide issu d'Asf1 et phosphorylé sur la thréonine 270 (Smolka et al., 2005). On ne peut cependant pas exclure que la thréonine 265 joue un rôle dans la liaison au domaine FHA1 de Rad53, même si les analyses bioinformatiques ne la placent pas parmi les bons candidats : elle est associée au motif pTxxE, sa probabilité d'être phosphorylée est très inférieure à celle de T270 (score = 0.26), et enfin elle est située dans une région moins conservée. Françoise Ochsenbein travaille actuellement à des études complémentaires qui permettront de déterminer précisément la thréonine reconnue.

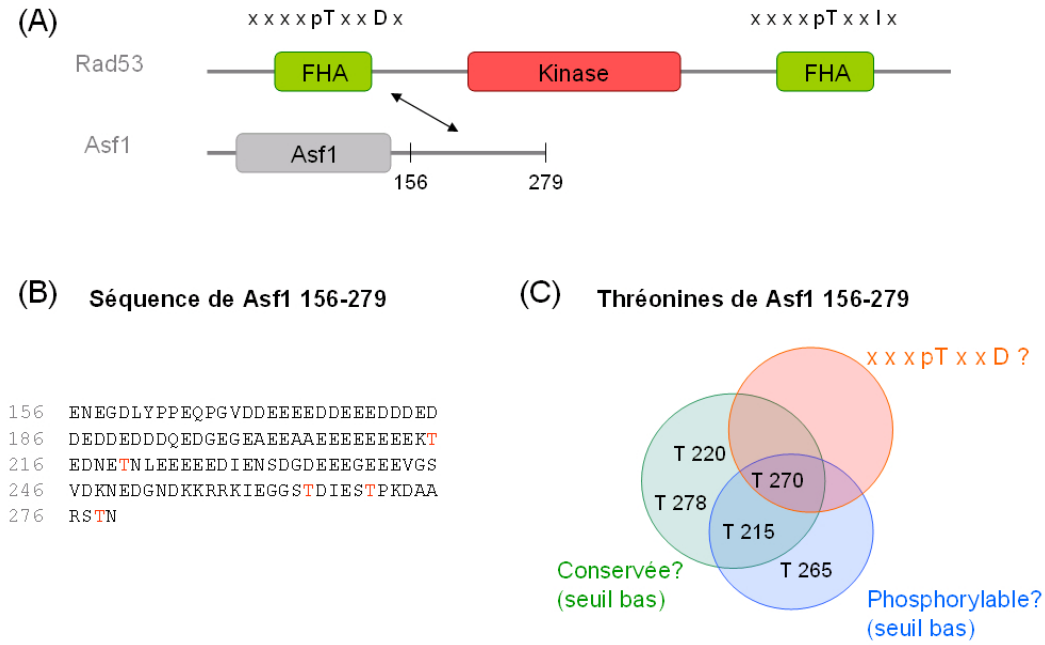


figure 43 : Etude de l'interaction entre le domaine FHA1 de Rad53 et Asf1. (A) Interaction entre le domaine FHA1 de Rad53 et le fragment 156-279 de Asf1. (B) Séquence du fragment 156-279 de la protéine Asf1. Les thréonines sont indiquées en rouge. (C) Analyse des différentes thréonines de Asf1 156-279 par les 3 critères de respect du motif de plus forte affinité du domaine FHA1 de Rad53, pTxxD (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 relativement à la thréonine (ensemble vert). Seule la thréonine T270 respecte ces trois critères simultanément avec les seuils abaissés (intersection des trois ensembles).

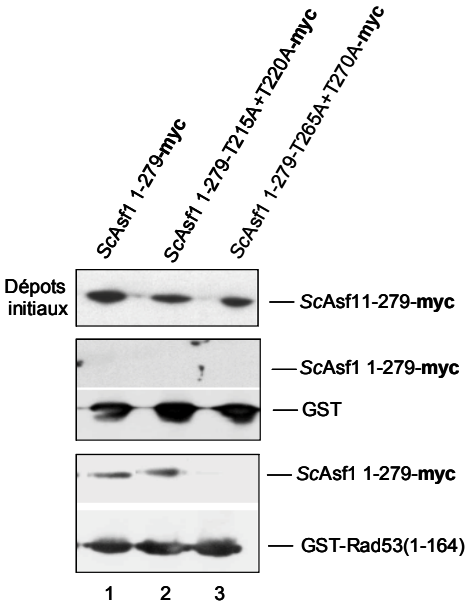


figure 44 : Western Blot des électrophorèses SDS-PAGE des échantillons de (puit 1) Asf1 1-279-myc ; (puit 2) Asf1 T215A+T220A-myc ; (puit 3) Asf1 T265A+T270A-myc. Le SDS-PAGE des dépôts initiaux (1 mg de protéines) des extraits cellulaires de levure est présenté sur le panneau du haut. Les SDS-PAGE correspondants aux trois échantillons incubés avec des quantités équivalentes (20 µg) de GST seule et de GST-Rad53 (1-164), sont présentés respectivement sur le panneau du milieu et sur le panneau du bas. Après différents lavages, on peut voir que les protéines Asf1 natives et mutées sur les thréonines 215 et 220 sont associées au domaine FHA1 de Rad53, alors que la protéine Asf1 mutée sur les thréonines 265 et 270 n'interagit plus avec ce dernier.

4.3 Application de cette méthode de détection à grande échelle dans le cadre du projet SplDER.

4.3.1 Recherche de partenaires des domaines FHA de Rad53 par double hybride.

Les succès rencontrés dans les prédictions précédentes nous ont conduit à explorer de façon plus systématique la robustesse de la stratégie bioinformatique.

Deux cribles double hybride à grande échelle ont été réalisés par Willy Aucher (post-doctorant au Laboratoire du Contrôle du Cycle Cellulaire, iBiTec-S) avec les domaines FHA1 et FHA2 de Rad53 comme proies. Pour ce qui concerne le domaine FHA1, les résultats sont présentés dans la **table 12** : ce crible a identifié 12 partenaires potentiels. Parmi ces protéines, seule Ptc2 est un partenaire connu du domaine FHA1 de Rad53. Deux protéines mises en évidence par ce crible sont impliquées dans le contrôle du cycle cellulaire et la réparation des dommages de l'ADN : Ptc2 et Cdc45. De façon identique à ce qui a été fait pour le domaine FHA1 de Rad53, un crible double hybride avec le domaine FHA2 de Rad53 comme proie a été réalisé. Les résultats sont présentés dans la **table 13**. Dans ce crible, on trouve un seul partenaire déjà connu de Rad53 : il s'agit de Dbf4.

Nom du partenaire potentiel	Nb de Hits	Code SwissProt	Code SGD
END3	16	P39013	YNL084C
CDC45 *	5	Q08032	YLR103C
VPS30	4	Q02948	YPL120W
ROG3 / YFJ2	4	P43602	YFR022W
YH02	3	P38887	YHR202W
YG18 *	2	P53207	YGR013W
PTC2 *	1	P39966	YER089C
TFC6 *	1	Q06339	YDR362C
YAP7 *	1	Q08182	YOL028C
PPZ1 *	1	P26570	YML016C
DCD1	1	P06773	YHR144C
GEA1	1	P47102	YJR031C

table 12 : Résultats du crible double hybride réalisé par Willy Aucher avec le domaine FHA1 de Rad53 comme proie. Dans la première colonne, les partenaires mis en évidences en fonction du processus cellulaire dans lequel ils interviennent : en vert, le cycle cellulaire et la réparation, en bleu, la transcription, en noir, les autres processus ou les protéines dont la fonction n'est pas connue. Les protéines nucléaires sont marquées d'une étoile. La seconde colonne indique le nombre de hits obtenus lors du crible double hybride. Les deux dernières colonnes indiquent respectivement les codes SwissProt et SGD (*Saccharomyces Genome Database*) de la séquence.

Nom du partenaire potentiel	Nb de Hits	Code SwissProt	Code SGD
STE5 *	11	P32917	YDR103W
NSE5 *	7	Q03718	YML023C
YB53	7	P38308	YBR203W
VPS72 *	4	Q03388	YDR485C
TRF5 *	3	P48561	YNL299W
YNR6 *	3	P53882	YNL176C
STD1 *	3	Q02794	YOR047C
ZDS2 *	3	P54786	YML109W
VAC17p	2	Q6QNF1	YCL063W
SSY1	1	Q03770	YDR160W
RPOM	1	P13433	YFL036W
MTH1	1	P35198	YDR277C
DBF4 *	1	P32325	YDR052C

table 13 : Résultats du crible double hybride réalisé par Willy Aucher avec le domaine FHA2 de Rad53 comme proie. Dans la première colonne, les partenaires mis en évidences en fonction du processus cellulaire dans lequel ils interviennent : en vert, le cycle cellulaire et la réparation, en bleu, la transcription, en orange, l'assemblage de la chromatine, en noir, les autres processus ou les protéines dont la fonction n'est pas connue. Les protéines nucléaires sont marquées d'une étoile. Les autres champs sont identiques à ceux de la table 12.

4.3.2 Détection du site reconnu par le domaine FHA1 de Rad53 sur Cdc45.

Cdc45 n'est pas un partenaire connu de Rad53. Nous avons choisi d'étudier cette interaction car c'est une protéine connue pour être nucléaire, comme Rad53. De plus, Cdc45 et Rad53 sont impliquées dans des processus cellulaires « voisins » : Cdc45 intervient dans l'initiation de la réplication de l'ADN (chez la levure, voir (Aparicio et al., 1999; Zou and Stillman, 2000) alors que Rad53 est un *checkpoint* du cycle cellulaire. Cette interaction est enfin l'une des plus fréquemment retrouvée dans le crible du domaine FHA1 de Rad53 (5 hits). Cdc45 étant un gène essentiel, une mutagenèse ciblée permettant d'abroger spécifiquement l'interaction entre Rad53 et Cdc45 serait une opportunité unique d'étudier le rôle de cette interaction et de ce gène.

Le fragment de Cdc45 interagissant avec le domaine FHA1 de Rad53 comporte 116 acides aminés, du résidu 154 au résidu 270 (**figure 45**-A). Il contient six thréonines indiquées en rouge sur la **figure 45**-B : T189, T195, T205, T245, T246 et T270. Aucune thréonine ne satisfait simultanément les trois critères lorsque les seuils les plus stringeants sont appliqués. Avec des seuils plus permissifs, la thréonine T189 est la seule à respecter les trois critères (**figure 45**-C). Elle respecte le motif pTxxD, son score de phosphorylabilité est de 0.33, et elle est

conservée dans 2/3 des séquences homologues chez les autres levures *Saccharomyces*. Nous avons donc proposé que T189 soit le site de Cdc45 reconnu par le domaine FHA1 de Rad53.

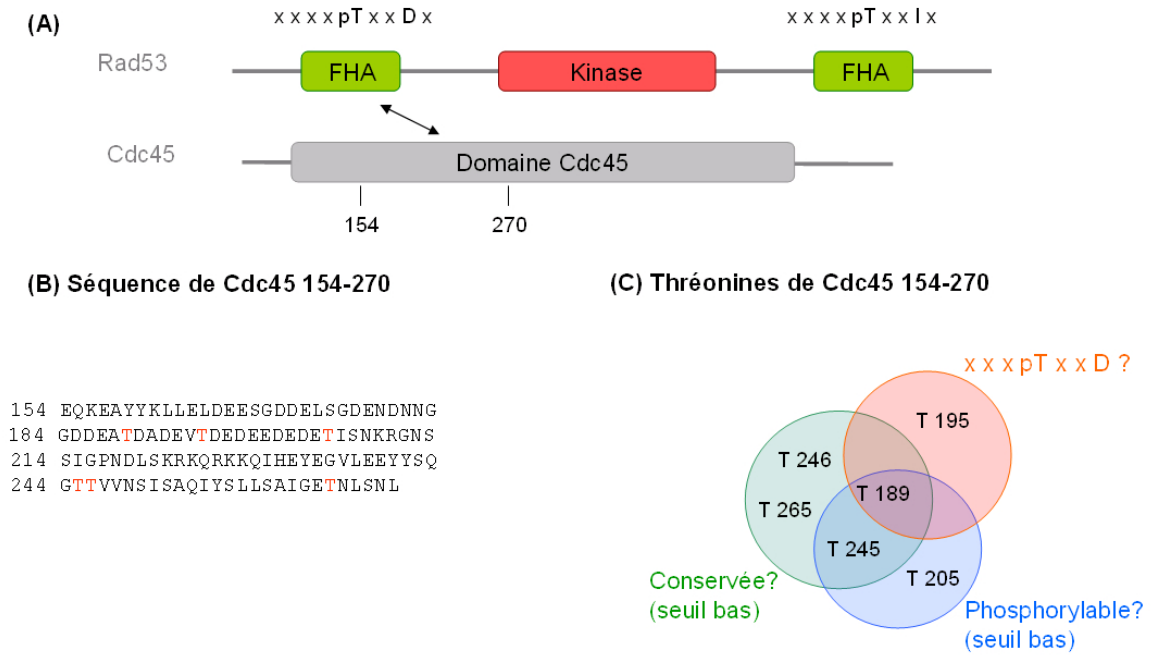


figure 45 : Etude de l'interaction entre le domaine FHA1 de Rad53 et Cdc45. (A) Interaction entre le domaine FHA1 de Rad53 et le fragment 154-270 de Cdc45 mise en évidence lors du crible double hybride. (B) Séquence du fragment 154-270 de la protéine Cdc45. Les thréonines sont indiquées en rouge. (C) Analyse des différentes thréonines de la séquence de Cdc45 154-270 par les 3 critères de respect du motif de plus forte affinité du domaine FHA1 de Rad53, pTxxD (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 (ensemble vert). Seule la thréonine T189 respecte ces trois critères simultanément, avec les seuils abaissés (intersection des trois ensembles).

Trois mutants de Cdc45 ont été produits : le mutant T189A d'une part, ainsi que les mutants T195A et T245A, qui serviront de contrôle. Les résultats de double hybride obtenus à partir de ces mutants (**figure 46**) montrent que la substitution T189A est la seule à affecter l'interaction. A l'heure actuelle l'étude de l'interaction entre le domaine FHA1 de Rad53 et Cdc45 se poursuit. Willy Aucher a récemment confirmé cette interaction par GST-pulldown (communication personnelle). Une étude du mutant Cdc45_T189A *in vivo* est également en cours.

Nous avons noté que le motif entourant T189 respecte les motifs consensus généralement phosphorylés par la caséine kinase 2 (CK2). Comme dans le cas de l'interaction Rad53-Ptc2, le rôle de cette kinase dans le processus d'activation des partenaires de Rad53 sera étudié.

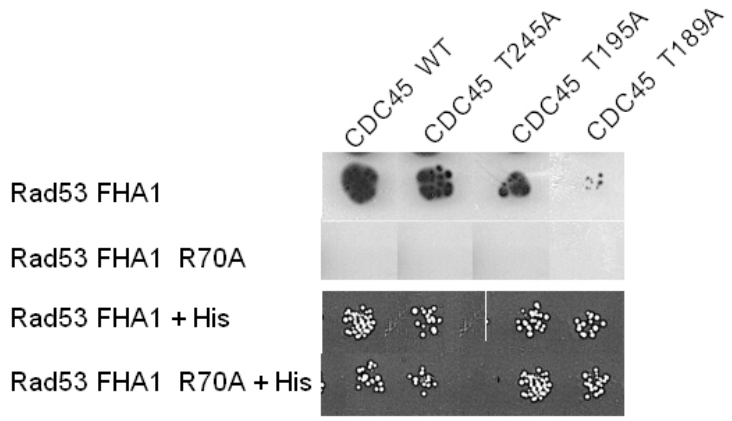


figure 46 : Etude par double hybride de l'interaction entre le domaine FHA1 de Rad53 et différents mutants de Cdc45 : CDC45 wild-type (WT), Cdc45_T245A, Cdc45_T195A et Cdc45_T189A. Le mutant Cdc45_T189A est affecté dans sa liaison au domaine FHA1 de Rad53. La substitution R70A du domaine FHA1 de Rad53 abroge l'interaction avec Cdc45 à la fois dans le wild-type et dans tous les mutants. Le panneau du bas atteste de la viabilité des colonies.

4.3.3 Détection du site reconnu par le domaine FHA1 de Rad53 sur Cdc7.

La kinase de régulation du cycle cellulaire Cdc7 est une sérine/thréonine kinase essentielle dont l'activité est nécessaire en phase S (Bousset and Diffley, 1998; Donaldson et al., 1998). Le complexe formé de Cdc7 et de sa sous-unité régulatrice Dbf4 interagit physiquement avec les origines de réplication (Dowell et al., 1994; Hardy, 1996; Hardy and Pautz, 1996). L'interaction mettant en jeu Rad53 et Cdc7 est connue (Kihara et al., 2000), néanmoins le site d'interaction précis n'a jamais été mis en évidence.

La séquence complète de Cdc7 comprend 24 thréonines (**figure 47**-B). Parmi celles-ci, 4 sont considérées comme phosphorylables par le programme NETPHOS-2.0 avec le seuil de détection le plus stringeant ; il s'agit de T2 (score = 0.67), T43 (score = 0.52), T239 (score = 0.55) et T484 (score = 0.88). Cette dernière thréonine, T484, est particulièrement intéressante car en plus du score élevé de phosphorylabilité, elle respecte le motif pTxxD particulièrement affin pour le domaine FHA1 de Rad53, et est parfaitement conservée dans l'alignement multiple construit à partir de Cdc7 et des séquences des protéines homologues chez les autres levures *Saccharomyces*. De plus, elle est la seule thréonine à respecter simultanément ces 3 critères (**figure 47**-C).

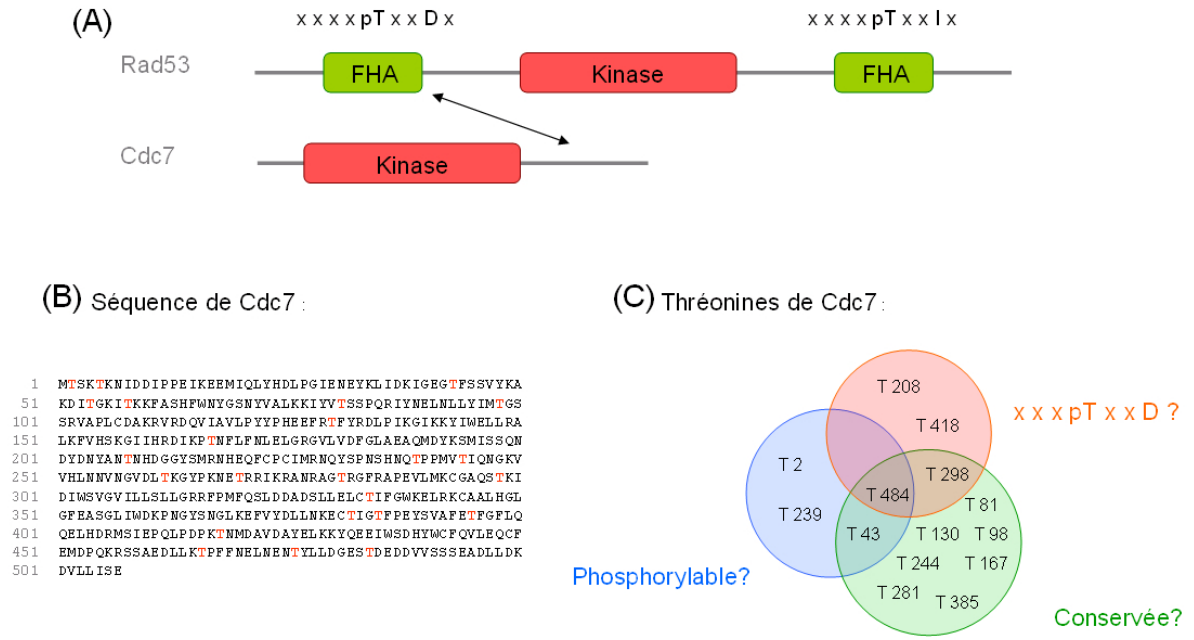


figure 47 : Etude de l'interaction entre le domaine FHA1 de Rad53 et Cdc7. (A) Interaction entre le domaine FHA1 de Rad53 et Cdc7. (B) Séquence de la protéine Cdc7. Les thréonines sont indiquées en rouge. (C) Analyse des différentes thréonines de la séquence de Cdc7 par les 3 critères de respect du motif de plus forte affinité du domaine FHA1 de Rad53, pTxxD (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 (ensemble vert). Seule la thréonine T484 respecte ces trois critères simultanément (intersection des trois ensembles).

En utilisant des critères moins stringeants pour la phosphorylabilité, la thréonine T298 parait elle aussi intéressante. Celle-ci respecte le motif pTxxD, et elle est fortement conservée tout comme l'aspartate en position +3 (dans 7 séquences sur 7). Son score de phosphorylabilité n'est que de 0.17, mais nous avons montré qu'un score relativement faible de 0.33 était affecté à la thréonine T189 de Cdc45 qui est celle reconnue par le domaine FHA1 de Rad53. Ainsi, nous avons également proposé que cette thréonine soit substituée en alanine pour servir de contrôle.

Willy Aucher a réalisé les deux mutants Cdc7_T298A et Cdc7_T484A et effectué des expériences de double hybride pour vérifier si ces deux mutants lient toujours le domaine FHA1 de Rad53. Les résultats obtenus (**figure 48**) montrent que le mutant Cdc7_T298A est capable d'interagir avec le domaine FHA1 de Rad53, tandis que l'interaction est abrogée avec le mutant Cdc7_T484A.

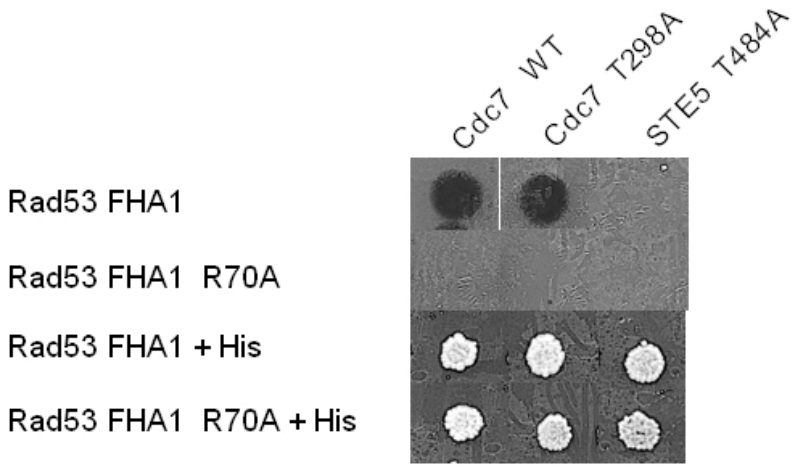


figure 48: Etude par double hybride de l'interaction entre le domaine FHA1 de Rad53 et différents mutants de Cdc7 : Cdc7 wild-type (WT), Cdc7 T298A et Cdc7 T484A. Le mutant Cdc7 T484A est affecté dans sa liaison au domaine FHA1 de Rad53. La substitution R70A du domaine FHA1 de Rad53 abroge l'interaction avec Cdc45 à la fois dans le wild-type et dans tous les mutants. La complémentation par HIS atteste que les colonies sont viables.

La thréonine T484 se trouve dans la région C-terminale de Cdc7, et on note que la séquence entourant cette thréonine, ^{480}DGESTDEDD^{488} est fortement acide. Encore une fois, nos analyses de séquences suggèrent que la kinase CK2 est responsable de la phosphorylation de la thréonine reconnue, comme pour Ptc2 et Cdc45. Cette hypothèse sera testée prochainement.

4.3.4 Détection du site reconnu par le domaine FHA2 de Rad53 sur STE5.

Le crible double hybride identifie Ste5 comme partenaire potentiel du domaine FHA2 de Rad53 (11 hits, c'est l'interaction la plus fréquente), alors que cette interaction n'est pas connue. Ste5 est une « *scaffolding protein* » (protéine d'échafaudage) ; son rôle est de recruter simultanément d'autres protéines afin de leur permettre d'interagir ensemble. Plus particulièrement, Ste5 participe au déclenchement de la conjugaison en réponse à l'α-facteur en formant un complexe avec les protéines Ste11, Ste7 et Fus3, ce qui permet à ces protéines d'activer la cascade de signalisation jusqu'au noyau (**figure 49**).

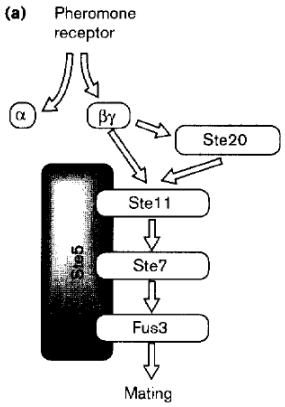


figure 49 : Représentation schématique du rôle de *scaffolding protein* de Ste5. Chez la levure, la voie de signalisation menant à la conjugaison utilise Ste5 comme *scaffolding protein* permettant de lier les membres du module formé de Ste11, Ste7 et Fus3. Figure extraite de (Garrington and Johnson, 1999).

L'identification d'une interaction entre Rad53 et Ste5 est assez inattendue puisque les deux protéines interviennent dans des processus cellulaires distincts et on peut envisager que cette interaction soit un « faux positif » fréquent dans la technique du double hybride. Nous avons néanmoins choisi de travailler sur Ste5 dans l'hypothèse qu'une interaction Rad53-Ste5 *in vivo* constituerait un point de couplage intéressant entre les voies de signalisation des dommages de l'ADN et le déclenchement de la conjugaison.

L'interaction constatée met en jeu le domaine FHA2 de Rad53 et le fragment protéique 712-845 provenant de STE5 (**figure 50**) qui contient dix thréonines. Parmi ces thréonines, la thréonine T809 est bien conservée chez les levures *Saccharomyces* (on la retrouve dans 6 séquences sur 7) et son score de phosphorylabilité est élevé (score=0.53). De plus, T809 est associée au motif de plus forte affinité, pTxxI. Puisque c'est la seule thréonine à respecter simultanément ces trois critères (avec les seuils abaissés), nous proposons que T809 soit la thréonine de Ste5 reconnue par le domaine FHA2 de Rad53.

Willy Aucher a étudié l'interaction observée par double hybride entre le domaine FHA2 de Rad53 et quatre mutants de Ste5 pour lesquels une thréonine du fragment interagissant a été substituée en alanine : T761A, T807A, T808A, qui serviront de contrôle, et T809A. La **figure 51** montre que la substitution T809A est la seule à abroger l'interaction. La procédure de sélection s'est donc à nouveau avérée efficace pour identifier la thréonine reconnue. Des études sont en cours pour valider l'interaction Ste5-Rad53 par GST-pulldown et pour caractériser le rôle de cette interaction potentielle *in vivo*.

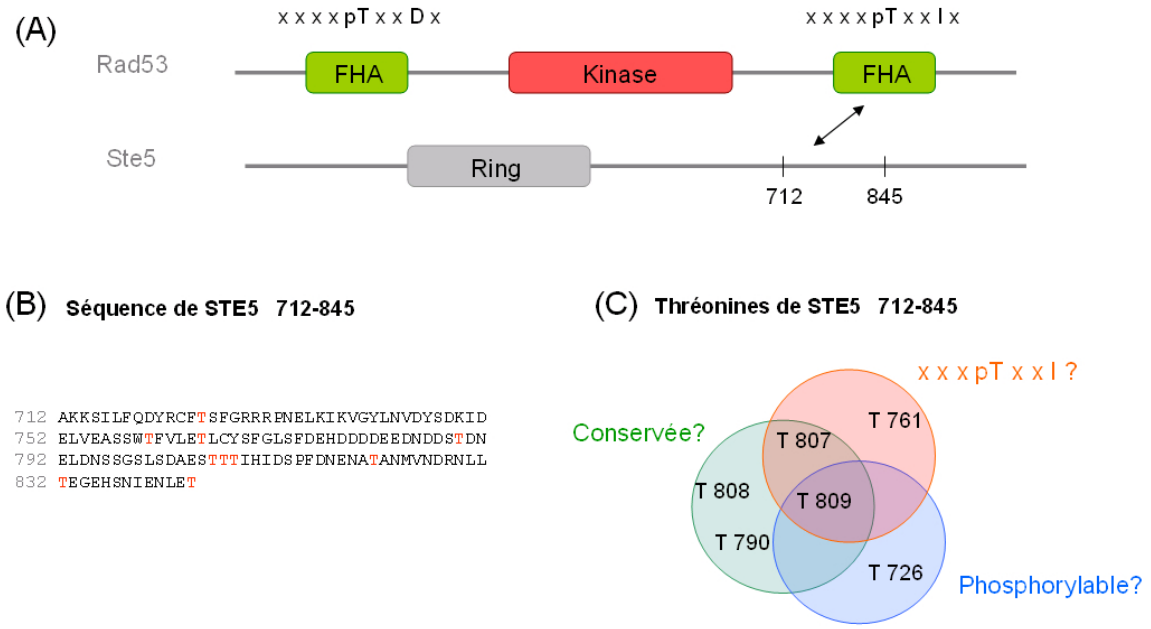


figure 50 : Etude de l'interaction entre le domaine FHA2 de Rad53 et Ste5. En haut : Interaction entre le domaine FHA2 de Rad53 et le fragment 712-845 de Ste5 mise en évidence lors du crible double hybride. A gauche : séquence du fragment 712-845 de la protéine Ste5. Les thréonines sont indiquées en rouge. A droite : analyse des différentes thréonines de la séquence de Ste5 712-845 par les 3 critères de respect du motif de plus forte affinité du domaine FHA2 de Rad53, pTxxI (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 relativement à la thréonine (ensemble vert). Seule la thréonine T809 respecte ces trois critères simultanément, avec les seuils abaissés (intersection des trois ensembles).

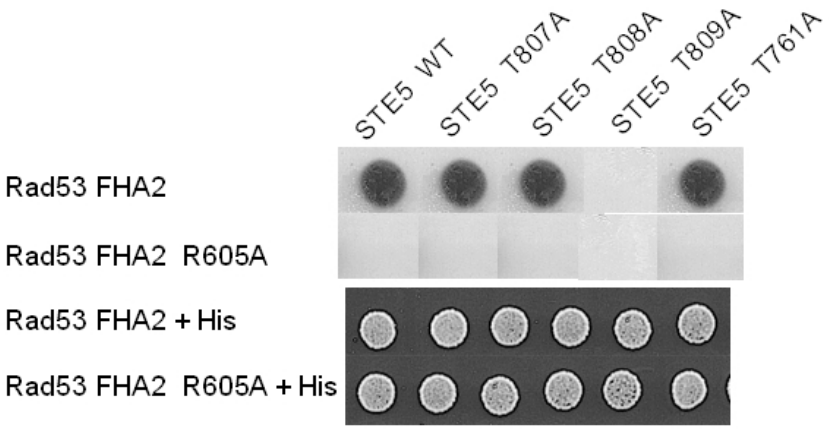


figure 51 : Etude par double hybride de l'interaction entre le domaine FHA2 de Rad53 et différents mutants de Ste5 : Ste5 wild-type (WT), Ste5 T807A, Ste5 T808A, Ste5 T809A et Ste5 T761A. Le mutant Ste5 T809A est affecté dans sa liaison au domaine FHA2 de Rad53. La substitution R605A du domaine FHA2 de Rad53 abroge l'interaction avec Ste5 à la fois dans le wild-type et dans tous les mutants, ce qui suggère que l'interaction est bien phospho-dépendante.

4.3.5 Etude de l'interaction entre le domaine FHA2 de Rad53 et NSE5

Une interaction entre le domaine FHA2 de Rad53 et la protéine NSE5 a été détectée par le crible double hybride de Willy Aucher. NSE5 est une sous unité essentielle du complexe Mms21-Smc5-Smc6, dont le rôle est crucial pour la réparation des dommages de l'ADN (Zhao and Blobel, 2005). L'étude de cette interaction entre Rad53 et NSE5 pourrait donc s'avérer très intéressante du point de vue de la compréhension des mécanismes des voies de surveillance et de réparation des dommages de l'ADN.

Le fragment de séquence NSE5 370-556 interagissant avec le domaine FHA2 de Rad53 comprend huit thréonines, dont aucune n'est parfaitement conservée : T373, T388, T404, T421, T429, T437, T464 et T499. Néanmoins, la thréonine et le motif qui l'entoure sont identiques dans plus de la moitié des séquences alignées pour toutes les thréonines du fragment à l'exception des deux thréonines C-terminales, T464 et T499. Les prédictions de phosphorylabilité mettent en avant T373, avec un score très élevé de 0.98, et également T421 (score = 0.54) et T437 (score = 0.65). Enfin, trois thréonines respectent le motif pTxxI/L : T373, qui est également bien conservée et prédite comme phosphorylable avec un score très élevé, T404, bien conservée mais prédite comme non phosphorylable, et T499, non conservée et prédite comme non phosphorylable (**figure 52**).

L'ensemble de ces résultats suggère que T373 est la meilleure candidate. Willy Aucher a donc produit des mutants NSE5_T373A, ainsi que des mutants NSE5_T404A qui serviront de contrôle, et étudier leur interaction avec le domaine FHA2 de Rad53 par double hybride. Les résultats, présentés sur la **figure 53** montrent qu'aucun des deux mutants de NSE5 n'abroge l'interaction, alors que nous nous attendions à ce que le mutant NSE5_T373A ne lie plus le domaine FHA2 de Rad53. Il est très intéressant de noter qu'on constate une interaction résiduelle entre FHA2 de Rad53 R605A et NSE5, à la fois sous sa forme native et mutée (**figure 53**). Comme nous l'avons expliqué dans le paragraphe traitant des aspects structuraux de Rad53 (4.1.1), le mutant R605A du domaine FHA2 de Rad53 est inefficace pour reconnaître les thréonines phosphorylées. L'ensemble de ces données suggère que le mode de reconnaissance de l'interaction entre le domaine FHA2 de Rad53 et la protéine NSE5 est particulier, et non phospho-dépendant.

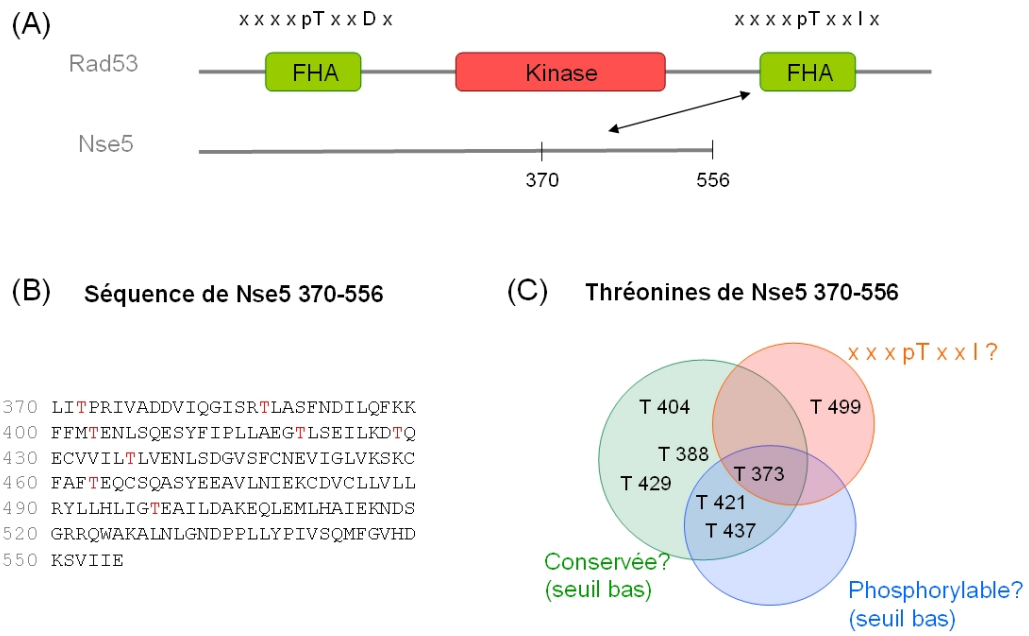


figure 52 : Etude de l'interaction entre le domaine FHA2 de Rad53 et Nse5. **(A)** Interaction entre le domaine FHA2 de Rad53 et le fragment hybride 370-556 de Nse5 mise en évidence lors du crible double hybride. **(B)** Séquence du fragment 370-556 de la protéine Nse5. Les thréonines sont indiquées en rouge. **(C)** Analyse des différentes thréonines de Nse5 370-556 par les 3 critères de respect du motif de plus forte affinité du domaine FHA2 de Rad53, pTxxI (ensemble rouge), de phosphorylabilité (ensemble bleu), de conservation de la thréonine et de l'acide aminé en position +3 (ensemble vert). Seule la thréonine T373 respecte ces trois critères simultanément, avec les seuils abaissés (intersection des trois ensembles).

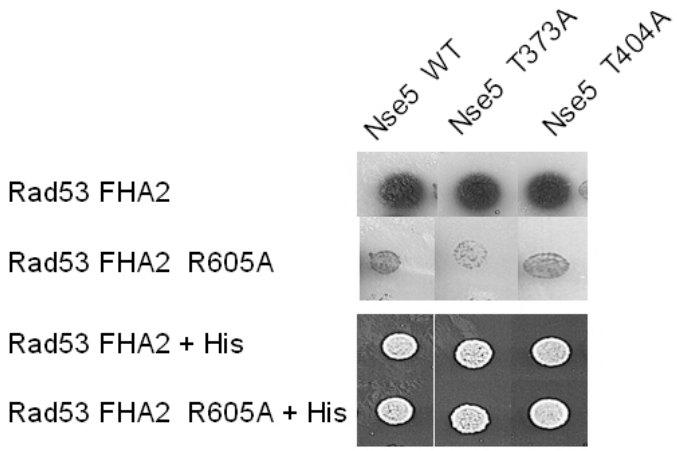


figure 53 : Etude par double hybride de l'interaction entre le domaine FHA2 de Rad53 et différents mutants de Nse5 : Nse5 wild-type (WT), Nse5_T373A, et Nse5_T404A. Aucun mutant de Nse5 n'est affecté dans sa liaison au domaine FHA2 de Rad53. De plus, on constate que la substitution R605A du domaine FHA2 de Rad53 n'abroge pas complètement l'interaction avec Nse5 ni dans le wild-type ni dans les deux mutants. Le panneau du bas atteste de la viabilité des colonies.

Des études complémentaires sont en cours et nous envisageons (i) de substituer l'une après l'autre toutes les thréonines du fragment afin d'exclure complètement le mode de reconnaissance phospho-dépendant, (ii) de confirmer cette interaction par GST-pulldown afin d'exclure la possibilité d'un faux positif de double hybride pour éventuellement (iii) étudier expérimentalement la structure du complexe *in vitro* par résonance magnétique nucléaire ou diffraction des rayons X.

4.4 Automatisation de l'analyse, mise en place du site web SpIDER.

4.4.1 Programme d'analyse de la séquence d'un partenaire.

L'analyse des partenaires est automatisée au sein d'un court programme écrit en Python qui effectue les interrogations des différents serveurs et les recoupements des résultats. Une page html au format du site internet du projet SpIDER est automatiquement générée. L'utilisation de ce programme nécessite 5 arguments décrits ci-dessous, et la **figure 54** illustre la sortie standard du programme:

```
$ ./analyze_sequence_of_FHA_partner.py
        [name]          ; le nom de la cible
        [code_SwissProt]        ; le code SwissProt de la cible
        [code_SGD]          ; le code SGD de la cible
        [FHAname]           ; le nom du FHA (sc_Rad53_FHA1 par ex)
        [pTxxD, pTxxI, pTxxL, pTxx[LI]]    ; le motif
```

Par cette procédure, nous avons analysé tous les partenaires des domaines FHA1 et FHA2 de Rad53 mis en évidence au cours du crible double hybride effectué par Willy Aucher. Marcus Smolka a récemment publié une étude (Smolka et coll, J. Cell Biol, accepté) au cours de laquelle il identifie 30 interactions entre les domaines FHA1 et FHA2 de Rad53 et d'autres protéines par un criblage protéomique ; les résultats de ce crible sont très différents des résultats connus et de ceux obtenus par le crible double hybride effectué par nos collaborateurs. Nous avons également analysé toutes ces interactions ; et les résultats seront disponibles sur le site du projet SpIDER (**Annexe B**).

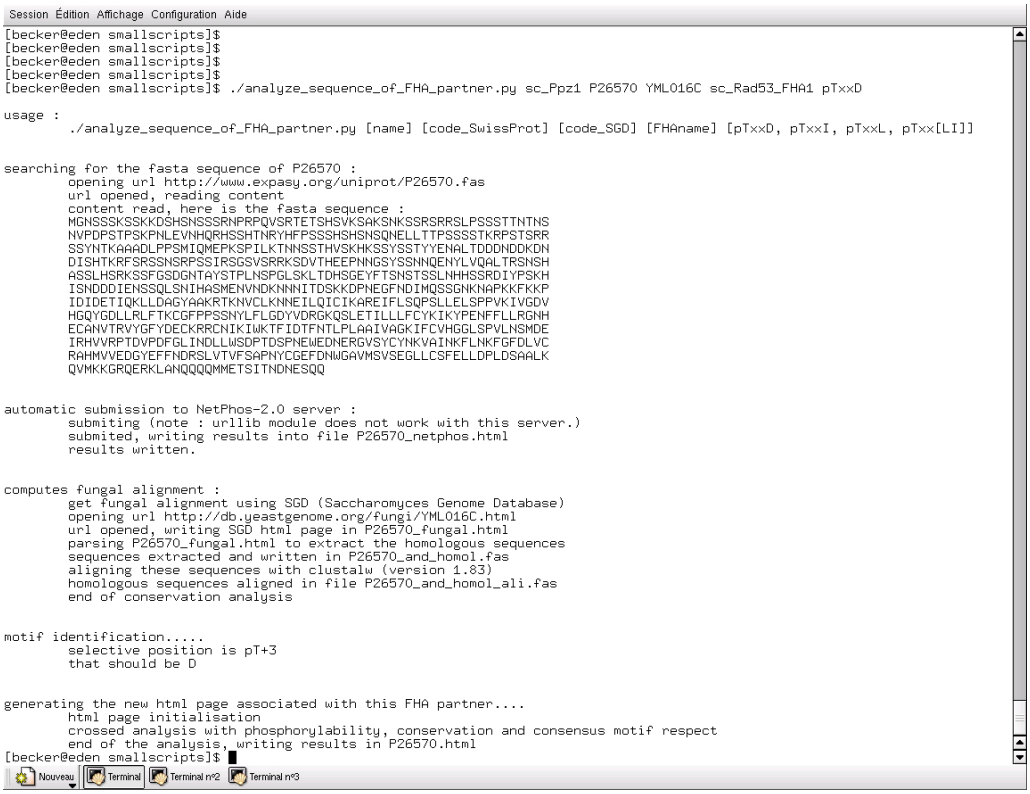


figure 54 : Sortie standard lors de l'analyse de l'interaction entre le domaine FHA1 de Rad53 et la protéine Ppz1. Le déroulement de l'analyse est indiqué au fur et à mesure de l'exécution : (1) recherche de la séquence protéique ; (2) analyse par NETPHOS-2.0 ; (3) récupération des séquences homologues, alignement, puis analyse de la conservation des thréonines ; (4) identification des motifs autour des thréonines ; (5) recoupement des informations et génération de la page web synthétisant les résultats.

4.4.2 Site du projet SpIDER : http://www-spider.cea.fr.

Pour centraliser tous les résultats d'analyse de séquence des partenaires, nous avons mis en place un site Internet qui sera disponible à l'adresse : http://www-spider.cea.fr. C'est un bon moyen de permettre aux différentes équipes impliquées dans le projet SpIDER de consulter les résultats de l'analyse de la séquence d'un partenaire quand elle le souhaite. La **figure 55** présente une page « standard » consacrée à l'étude de la séquence d'un partenaire. Cette page se décompose en deux sections.

La première partie regroupe une série de liens utiles pour analyser la séquence : (i) la séquence au format fasta ; (ii) les pages SWISSPROT, SGD et PFAM correspondantes ; (iii) un alignement de séquences avec les homologues proches ; et (iv) le fichier de sortie généré par NETPHOS-2.0 sur la phosphorylabilité des thréonines. Dans la seconde partie, chaque

thréonine de la séquence du partenaire est analysée et les résultats sont synthétisés dans un tableau. Pour chaque thréonine, la position et la séquence environnante sont indiquées, ainsi que le motif (pTxxD, pTxxI,…) et les scores exacts de phosphorylabilité et de conservation. Un code de couleur permet de mettre en évidence le respect ou non des critères de phosphorylabilité, de conservation et de respect du motif spécifique.

A l'heure actuelle, l'accès au site du projet SpIDER est protégé car les résultats (notamment ceux du crible double hybride) ne sont pas encore publiés. Par la suite, le serveur sera ouvert en accès libre.

INFORMATIONS ABOUT THE COMPLETE SEQUENCE OF SC_CDC3:

SwissProt entry - here - and fasta format sequence - here -
Search of domains within the sequence with pfam - here -
SGD entry- here - Close Homologs (fungis) - fasta - fasta format
Probability that the threonines within the sequence are phosphorylated - here -

Informations sur la séquence et liens divers

TABLE SUMMERIZING POTENTIAL INTERACTING SITES

pos	seq	pThr ?	motif ?	T cons ?	+3 cons ?
44	DSQYTNGTQ	0.014	pTxxT	(8/9)	(8/9)
47	YTNGTQNDD	0.421	pTxxD	(8/9)	(4/9)
73	GMGITSSQS	0.131	pTxxQ	(6/9)	(5/9)
133	GIGKTTLMK	0.027	pTxxM	(7/9)	(7/9)
134	IGKTTLMKT	0.839	pTxxK	(9/9)	(7/9)
138	TLMKTLFNN	0.017	pTxxN	(8/9)	(8/9)
205	NVIDTEGFG	0.087	pTxxF	(8/9)	(8/9)
260	FIEPTGHYL	0.021	pTxxY	(9/9)	(7/9)
292	SDILTDEEI	0.091	pTxxE	(9/9)	(9/9)
302	SFKKTIMNQ	0.061	pTxxN	(3/9)	(6/9)
396	LKERTSKIL	0.570	pTxxI	(9/9)	(2/9)
435	LEEKTLHEA	0.869	pTxxE	(6/9)	(8/9)
450	IEMKTVFQQ	0.083	pTxxQ	(7/9)	(8/9)
468	QKSETELFA	0.011	pTxxF	(6/9)	(8/9)
482	KEKLTKQLK	0.071	pTxxL	(7/9)	(8/9)
513	SPVPTKKKG	0.895	pTxxK	(7/9)	(8/9)

Résultats de l'analyse de toutes les thréonines contenues dans la séquence

figure 55: Extrait du site dédié au projet SpIDER. Cette page étudie la séquence de la protéine Cdc3 dont l'interaction avec le domaine FHA1 de Rad53 a été mise en évidence lors du crible double hybride réalisé par Willy Aucher. La partie haute rassemble les liens utiles pour étudier la séquence de Cdc3. En bas, le tableau rassemble tous les résultats de l'analyse des thréonines de Cdc3. Les deux premières colonnes listent pour chaque thréonine sa position absolue dans la séquence et la séquence qui l'entoure (4 résidus de part et d'autre). La troisième colonne donne le score exact de phosphorylabilité attribué par NETPHOS-2.0. La quatrième colonne extrait le motif associé à cette thréonine. Les deux dernières colonnes indiquent la conservation de la thréonine et de l'acide aminé en position +3 dans la séquence des autres levures *Saccharomyces*. Le code des couleurs est : fushia lorsque le critère est respecté avec les seuils les plus stringeants, rose pâle si le critère n'est respecté qu'avec des seuils plus permissifs, et vert lorsque le critère n'est pas respecté.

4.5 Discussion

4.5.1 Importance de la prise en compte simultanée des trois critères.

Au regard des différentes interactions mettant en jeu des domaines FHA de Rad53 sur lesquelles la stratégie décrite en 4.2.1 a été appliquée, il est remarquable de noter que les trois paramètres pris en compte sont complémentaires.

Dans le cas de l'interaction entre le domaine FHA1 de Rad53 et Ptc2, il n'aurait pas été possible d'identifier de manière unique la thréonine T376 sans tenir compte du critère de respect du motif de plus forte affinité, pTxxD. Si ce critère n'avait pas été utilisé, la thréonine T48 aurait elle aussi été une candidate potentielle car elle est à la fois prédite comme phosphorylable et parfaitement conservée au sein des espèces proches. L'interaction entre Ste5 et le domaine FHA2 de Rad53 souligne quant à elle l'importance du critère de phosphorylabilité. Dans cet exemple, si la probabilité de phosphorylation n'avait pas fait partie des critères évalués, il n'aurait pas été possible de discriminer entre la thréonine T809, et la thréonine T807. La conservation du site reconnu par le domaine FHA dans les séquences des autres levures *Saccharomyces* semble jouer un rôle moins important. Il est en effet possible de constater qu'au sein des cinq exemples d'interactions présentés pour lesquels la thréonine reconnue a été identifiée par notre analyse bioinformatique, le critère de conservation n'était pas indispensable. Cette constatation peut s'expliquer, car nous avons restreint la recherche d'homologues aux seules séquences des autres levures *Saccharomyces,* ce qui fait que l'identité de séquence au sein des alignements multiples générés est élevée et que la plupart des thréonines sont conservées. Ce critère n'est donc pas discriminant dans ce cadre précis.

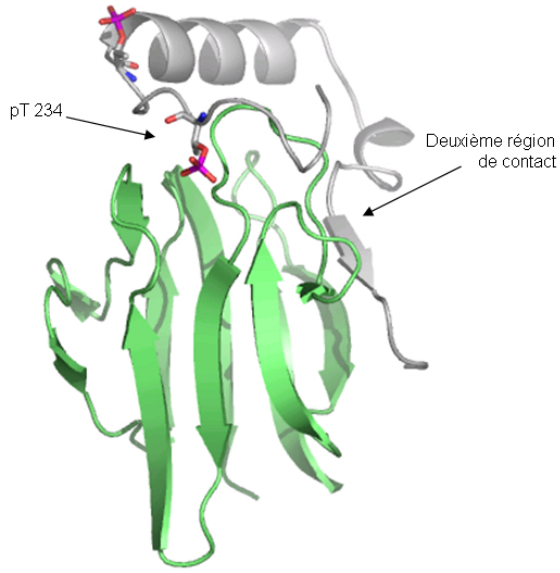
Une étude récente de l'équipe de Colin Watts (*Garvan Institute of Medical Research*, Australia) a mis en évidence une interaction entre le domaine FHA de la protéine humaine Chk2 et la protéine EDD, dont la fonction cellulaire précise n'est pas établie (Henderson et al., 2006). Ces travaux ont notamment tenté de détecter le site de EDD reconnu par le domaine FHA de Chk2 en identifiant toutes les thréonines de EDD respectant le motif pTxxI, mis en évidence par Daniel Durocher comme le motif le plus affin pour ce domaine FHA (Li et al., 2002). Après avoir identifié les 15 thréonines respectant le motif TxxI, ils ont choisi de substituer quatre d'entre elles en alanines, soit individuellement, soit simultanément. Les résultats

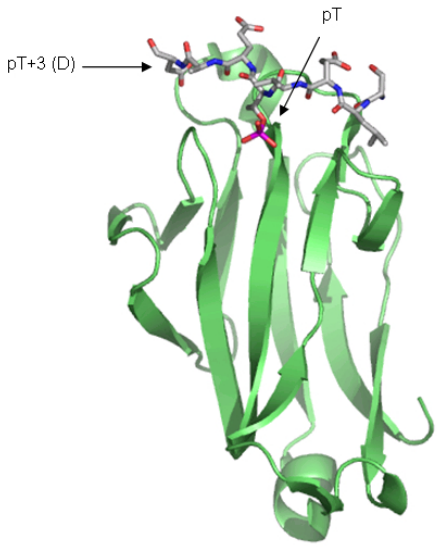
obtenus ne leur ont pas permis de mettre en évidence une thréonine dont la substitution en alanine abroge l'interaction avec le domaine FHA de Chk2. Ces résultats négatifs soulignent qu'il est important de tenir compte d'autres critères que celui du respect du motif de plus haute affinité afin de limiter le nombre de thréonines candidates et donc le nombre de mutants à étudier.

4.5.2 Modes de liaisons des domaines FHA

Peu après que les domaines FHA ont été mis en évidence en tant que domaines médiateurs d'interactions protéine-protéine reconnaissant des phospho-thréonines en 1999 (Durocher et al., 1999), les premières expériences de criblage de bibliothèques de phospho-peptides ont été réalisées afin de mettre en évidence une sélectivité particulière de ces domaines pour un motif protéique. Ces études ont mis en évidence une sélectivité en position pT+3, très versatile d'un domaine FHA à l'autre (Byeon et al., 2001; Durocher et al., 1999; Durocher et al., 2000; Li et al., 2002; Liao et al., 2000). Ainsi, le mode de reconnaissance proposé en 2002 par Daniel Durocher et Peter Jackson est que les domaines FHA se lient à des motifs linéaires courts, contenant des phospho-thréonines, et respectant un motif particulièrement affin propre à chaque domaine FHA, la position pT+3 étant responsable d'une grande partie de la sélectivité (Durocher and Jackson, 2002). L'interaction entre le domaine FHA2 de Rad53 et Rad9, mise en évidence en 1999 (Byeon et al., 2001; Durocher et al., 1999) et dans laquelle la thréonine de Rad9 reconnue par le domaine FHA2 de Rad53 respecte le motif de plus forte affinité mis en évidence lors des criblages de bibliothèques de phospho-peptides (Durocher et al., 2000), suggère que ce mode de reconnaissance est réel *in vivo*.

Cependant, quelques articles publiés par la suite ont remis en question ce modèle de reconnaissance en mettant en évidence des exceptions paradoxales. Il a notamment été établi par des expériences biochimiques et la publication d'une structure du domaine FHA de Ki67 en complexe avec un fragment de la protéine hNifk (Byeon et al., 2005; Li et al., 2004), que ce domaine FHA pouvait reconnaître des motifs contenant des phospho-thréonines plus longs (plus de 40 acides aminés), et sans sélectivité particulière sur la position pT+3. La **figure 56** présente la structure de ce complexe (**figure 56**-A), que l'on peut mettre en regard de la structure du complexe FHA1 de Rad53 avec un peptide court (**figure 56**-B).





(A) **Reconnaissance de fragments longs** –
Structure du Complexe formé par le domaine FHA de Ki67 et un fragment de 44 résidus issu de hNifk.

(B) **Reconnaissance de fragments courts** –
Structure du Complexe formé par le domaine FHA1 de Rad53 et un peptide respectant le motif pTxxD.

figure 56 : Structures de complexes mettant en jeu des domaines FHA et introduisant les deux modes de liaisons des domaines FHA identifiés à ce jour. (A) Le modèle « reconnaissance de fragments longs » représenté par la structure du complexe entre le domaine FHA de Ki67 et un fragment protéique de 44 résidus provenant de hNifk. La structure a été déterminée par RMN (code PDB 2AFF, modèle 1) : le domaine FHA est en ruban vert, le fragment issu de hNifk est en ruban blanc, les deux thréonines phosphorylées de ce fragments pT234 et pT238 sont identifiées. C'est la thréonine pT234 qui est reconnue par la poche de reconnaissance conservée du domaine FHA de Ki67. (B) Le modèle « reconnaissance de fragments courts », illustré par la structure du domaine FHA1 de Rad53 en complexe avec un fragment court respectant le motif déterminé par criblage de bibliothèque de phospho-peptides comme le motif le plus affin : pTxxD. La structure a été résolue par diffraction des rayons X (code PDB 1G6G). Le domaine FHA est en ruban vert et le peptide phosphorylée en gris.

Bien que la poche de reconnaissance de la thréonine phosphorylée soit identique dans les deux complexes, on constate de nettes différences. Dans le complexe entre le domaine FHA de Ki67 et le fragment de hNifk, on constate la formation d'un feuillet mettant en jeu un brin de l'extrémité hybride du fragment de hNifk et le brin β_4 du domaine FHA de Ki67. Ces travaux démontrent qu'il existe pour les domaines FHA un autre mode de reconnaissance, qui privilégie des fragments plus longs et sans sélectivité particulière concernant la position pT+3.

Enfin, une étude réalisée par Brietta Pike en 2004 et caractérisant l'interaction entre le domaine FHA1 de Rad53 et la protéine Mdt1 met en évidence que la thréonine reconnue par le domaine FHA1 de Rad53 est associée à un motif pTxxI, alors que le motif de plus forte affinité du domaine FHA1 de Rad53 privilégie un acide aminé acide en position pT+3 (Pike

et al., 2004). Le même paradoxe a été montré dans le cadre de l'interaction entre le domaine FHA1 de Rad53 et Rad9 (Schwartz et al., 2002).

La conjugaison de ces deux éléments :

(i) la résolution d'une structure de complexe attestant d'un mode de reconnaissance où le fragment reconnu est plus long et ne fait pas intervenir de sélectivité en position pT+3 et ;

(ii) la mise en évidence de deux interactions entre un domaine FHA et un fragment protéique ne respectant pas son motif de plus haute affinité ;

a conduit à une remise en question du premier modèle de reconnaissance de peptide courts par les domaines FHA (Mahajan et al., 2005). Peu d'exemples témoignent en effet de la pertinence du modèle basé sur la reconnaissance de peptides courts : avant les travaux présentés dans ce chapitre, seule l'interaction entre le domaine FHA2 de Rad53 et Rad9 a montré que la thréonine reconnue respecte la règle du motif pT+3.

Nos résultats permettent d'apporter de nouveaux éléments à cette discussion. En effet, pour sélectionner les thréonines reconnues par les domaines FHA1 et FHA2 de Rad53, nous avons utilisé le respect des motifs courts de plus haute affinité, déterminés par les criblages de bibliothèques de peptides, comme des critères de sélection. Les cinq exemples d'interactions mettant en jeu un domaine FHA de Rad53 que nous avons présentés et pour lesquels nous avons effectivement détecté la thréonine reconnue, sont autant d'exemples allant dans le sens du modèle de reconnaissance initialement présenté, puisque la nature de l'acide aminé en position +3 est un critère de sélection pertinent. Nos résultats suggèrent ainsi que les deux modes de reconnaissance « fragment long » et « fragment court » co-existent *in vivo*.

Chaque domaine FHA est donc susceptible de présenter deux modes d'interaction. Si dans certains cas, notre analyse basée sur l'hypothèse « fragment court » échoue, il est possible d'imaginer un mode de liaison alternatif que seules des analyses structurales poussées pourront caractériser. Nos résultats suggèrent que cela pourrait être le cas de l'interaction entre le domaine FHA1 de Rad53 et NSE5.

Chapitre 5 : Prédiction des spécificités de reconnaissance des PRMs : criblage *in silico* sur squelette rigide

Au cours du chapitre précédent, nous avons vu que la connaissance du motif spécifiquement reconnu par un PRM constitue un facteur clé dans le processus d'identification du site de liaison. Au cours de ce chapitre, nous abordons la question de la prédiction des spécificités d'interaction. Cette prédiction est particulièrement délicate car elle suppose non seulement de reconnaître le motif le plus favorable mais également d'exclure l'ensemble des solutions alternatives. Pour aborder cette question complexe, nous avons dans un premier temps restreint l'étude à des cas où le squelette peptidique de la structure du complexe domaine/peptide est maintenu rigide. La complexité supplémentaire liée à l'introduction de mouvements du squelette sera l'objet du chapitre 6.

Dans le chapitre 5, nous explorons comment la prise en compte du mouvement restreint aux chaînes latérales et l'estimation des énergies d'interaction permet de prédire les spécificités de liaison des PRMs. Cette question s'apparente au problème de design d'une séquence sur le squelette rigide d'une protéine ou d'un complexe. Pour cette raison, nous avons exploité les stratégies développées pour le design des protéines afin d'évaluer leurs performances à prédire les motifs spécifiquement reconnus. Deux familles de domaines médiateurs d'interactions au cœur de cette thèse ont retenu notre attention : les domaines FHA et les tandems de domaines BRCT.

5.1

Introduction

5.1.1 Objectif : étudier la faisabilité de criblage de PRMs *in silico*.

Au sein des différentes voies de signalisation, un grand nombre d'interactions essentielles sont médiées par des PRMs (Pawson and Nash, 2003). Une compréhension précise des mécanismes de transduction du signal nécessite donc de mettre en évidence les bases moléculaires responsables de la spécificité de reconnaissance de ces interactions transitoires, souvent associées à des phosphorylations ou à d'autres modifications post-traductionnelles.

Les principales méthodes utilisées pour déterminer la spécificité de reconnaissance sont des criblages de bibliothèques de peptides soit *in vivo* (Tong et al., 2002), soit *in vitro* (Li et al., 2002; Vetter and Zhang, 2002). Ces deux approches présentent des inconvénients. Les méthodes *in vivo*, qui utilisent les techniques de *phage-display*, ne permettent d'étudier que les PRMs qui lient des peptides non associés à des modifications post-traductionnelles. D'un autre côté, les méthodes de criblage *in vitro* sont très onéreuses et nécessitent un travail expérimental conséquent.

Dans ce contexte, nous avons étudié la faisabilité d'une approche de criblage *in silico* qui permette de prédire le motif spécifiquement reconnu par un PRM sur la base de sa structure tridimensionnelle.

5.1.2 Choix des familles de PRMs qui serviront de modèle d'étude.

Deux familles de domaines médiateurs d'interactions protéine-protéine ont retenu notre attention : les domaines FHA et les tandems de domaines BRCT. Ces deux PRMs sont associés à des motifs contenant des résidus phosphorylés : phospho-thréonine pour les domaines FHA et phospho-sérine pour les tandems de domaines BRCT.

Le choix de ces deux familles comme modèles d'étude repose sur (i) l'intérêt méthodologique qu'elles mettent en jeu ; (ii) l'intérêt biologique des protéines qui contiennent ces domaines ; et enfin sur (iii) l'existence de données expérimentales essentielles pour valider les tests *in silico*. La **table 14** explicite ces différents aspects.

	Domaines FHA	**Tandems BRCT**
Intérêt Méthodologique	La surface d'interaction entre le domaine FHA et le peptide est limitée et bien définie. La zone du domaine FHA qui est en contact avec le peptide est une région hautement flexible car elle met en jeu des boucles séparant des éléments de structure secondaire conservés. C'est donc un cas de figure particulièrement délicat. La sélectivité est très versatile et repose principalement sur une position qui peut être soit pT+3 soit pT-3. Aucune méthode basée sur l'analyse des séquences ne permet de prédire la sélectivité des domaines FHA.	Dans les exemples mis en évidence, la sélectivité repose sur une seule position (pS+3). Aucune méthode basée sur l'analyse des séquences ne permet de prédire la sélectivité des tandems de domaines BRCT.
Importance Biologique	Régulation de l'activité de certaines protéines clés de voies de surveillance des dommages de l'ADN. Le motif spécifiquement reconnu reste inconnu pour la plupart des domaines FHA.	Régulation de protéines clés des voies de surveillance des dommages de l'ADN. Les domaines BRCT en tandems sont peu étudiés à ce jour et le motif spécifiquement reconnu n'est déterminé que pour deux tandems de domaines BRCT.
Données Disponibles	Le domaine FHA1 de la protéine de levure Rad53, de structure déterminée par diffraction des rayons X (code 1G6G), reconnaît spécifiquement les motifs pTxxD ; Le domaine FHA de la protéine humaine Chk2, de structure déterminée par diffraction des rayons X (code 1GXC), reconnaît spécifiquement les motifs pTxxL/I ; Le domaine FHA de la polynucléotide kinase (Pnk) de mammifère, de structure résolue par diffraction des rayons X (code 1YJM) qui se lie à un motif DxxpT.	Le tandem de domaines BRCT de la protéine humaine Brca1, impliquée notamment dans le cancer du sein, possède une structure résolue par diffraction des rayons X (code 1T15) et reconnaît les motifs protéiques pSxxY/F. Le tandem de domaines BRCT de la protéine humaine Mdc1, de structure expérimentale résolue par diffraction des rayons X (code 2AZM), reconnaît les motifs pSxxY/F.

table 14 : Table présentant les différents points qui nous ont menés à choisir les domaines FHA et les tandems de domaines BRCT comme systèmes d'étude. Trois facteurs sont pris en compte : (i) l'intérêt méthodologique que constituent ces familles de PRMs ; (ii) l'intérêt biologique, c'est-à-dire le rôle majeur des protéines contenant ces PRMs dans des processus cellulaires très étudiés ; et (iii) la disponibilité de données expérimentales permettant d'effectuer et de valider les prédictions *in silico*.

On dispose de données expérimentales incluant les structures de très haute résolution et les résultats de criblages *in vitro* de bibliothèques de peptides pour au moins trois domaines

FHA et deux tandems BRCT : (1) le domaine FHA de la protéine humaine Chk2, (2) le domaine FHA N-terminal de la protéine de levure Rad53, (3) le domaine FHA de la polynucléotide kinase de mammifère, (4) le tandem BRCT de Brca1 et (5) le tandem BRCT de Mdc1. Pour les domaines FHA comme pour les tandems BRCT, aucune méthode basée sur l'analyse de séquence ne permet de prédire le motif spécifiquement reconnu par ces domaines. Pour cette raison, nous avons exploré si une stratégie de criblage *in silico* sur la base de la structure des domaines permet de retrouver les résultats des criblages expérimentaux.

5.1.3 Processus de criblage *in silico*.

Pour prédire le motif spécifiquement reconnu par chacun de ces domaines, nous utilisons leur structure complexée comme base structurale et remplaçons le ligand peptidique par une séquence poly-alanine sur toutes les positions sauf celle portant la modification post-traductionnelle. Le principe du criblage *in silico* consiste alors à explorer, à chaque position du peptide, les 20 acides aminés en cherchant par des mouvements adaptés des chaînes latérales à optimiser les interactions entre l'acide aminé muté et la surface du domaine FHA ou BRCT. Cette procédure s'apparente au *design* et à l'ingénierie des séquences visant à optimiser la stabilité d'une structure donnée. Pour cette raison, nous avons dans un premier temps testé la capacité des méthodes actuelles de *design* à prédire la séquence la plus affine pour un domaine FHA ou un tandem de domaines BRCT donné.

Le principe de la méthode est illustré sur la **figure 57** dans le cas du domaine FHA de Chk2. Dans un premier temps, le peptide complexé et les résidus du domaine FHA en contact sont élagués de leurs chaînes latérales. Pour chaque position du peptide et pour chaque acide aminé, on modélise l'interface la plus favorable. Enfin, chaque modèle est évalué à l'aide d'une fonction d'évaluation estimant l'énergie libre du complexe, et les acides aminés stabilisant l'interface sont identifiés.

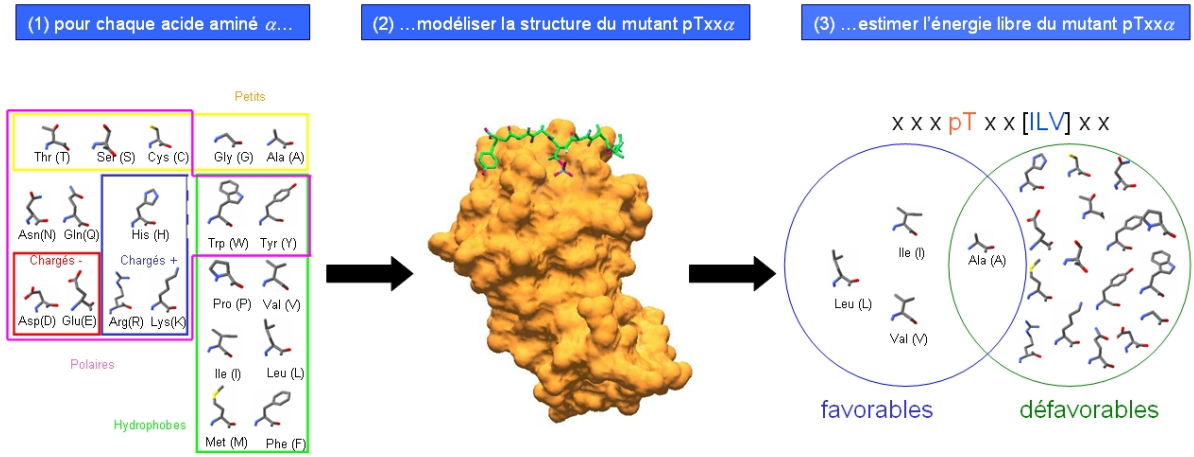


figure 57 : Illustration du procédé de criblage *in silico*. Pour chaque acide aminé a, on modélise la structure d'un mutant où la position spécifique est substituée par a : il s'agit donc d'un problème de modélisation. Dans la figure ci-dessus, la position pT+3 a été substituée par une tyrosine. Ensuite, dans une seconde étape, les différents mutants sont évalués et on peut détecter quels acides aminés sont favorables/défavorables lorsqu'ils sont placés à la position responsable de la sélectivité. Dans l'exemple, les acides aminés L/I et V sont les plus favorables.

Cette problématique soulève deux questions importantes.

- Lorsque l'acide aminé sélectionné α est effectivement celui reconnu *in vitro* / *in vivo*, l'interface modélisée est-elle proche de l'interface native du complexe ?
- Les fonctions de score couramment utilisées pour le *design* sont-elles suffisamment précise pour approximer l'énergie libre des différents complexes modélisés et différencier celui reproduisant l'interface native ?

Deux stratégies utilisant deux fonctions d'évaluation différentes ont été testées : (i) la fonction de score ROSETTADESIGN (que nous nommerons simplement ROSETTA dans la suite du manusrit), qui possède son propre système d'exploration conformationnelle ; et (ii) la fonction FOLDEF qui permet d'approximer l'énergie libre d'un complexe mais n'est pas couplé à une fonction de recherche conformationnelle. Le principe de fonctionnement de ces deux approches a été décrit dans l'introduction générale et leur champ d'application est rappelé **figure 58**. Puisque le programme FOLDEF estime l'énergie d'une structure sans permettre d'introduire des mouvements dans les chaînes latérales, nous avons recherché un programme permettant de modéliser la structure des mutants en amont de l'évaluation par FOLDEF. Les programmes de modélisation et d'optimisation du placement des chaînes latérales sur un squelette fixe, présentés dans l'introduction (chapitre 1), sont adaptés à ce

problème d'exploration conformationnelle. Néanmoins, les publications qui présentent les performances de ces algorithmes utilisent des bases de tests différentes ce qui ne permet pas l'identification objective de la meilleure méthode. Nous avons donc dans un premier temps évalué les capacités prédictives de quatre programmes récents de prédiction de la conformation des chaînes latérales sur un même ensemble de structures tests.

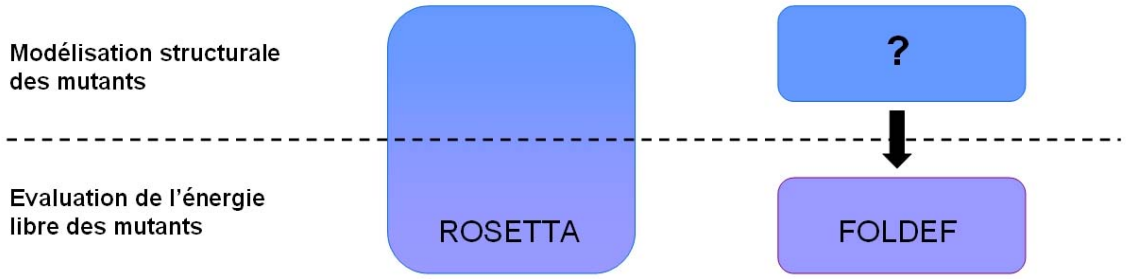


figure 58 : Deux méthodes de criblage *in silico*. La première méthode utilise la fonction de score ROSETTA, et possède son propre moteur de recherche pour la modélisation des mutants. La seconde approche utilise la fonction de score FOLDEF. Cette fonction n'est pas couplée à un moteur d'exploration conformationnelle pour la modélisation des mutants. Il faut donc coupler FOLDEF à un programme permettant de réaliser cette opération.

5.2 Optimisation du placement des chaînes latérales sur un squelette fixe.

5.2.1 Etude préliminaire : comparaison de différentes heuristiques existantes.

L'étude comparative de quatre programmes de prédiction de la conformation des chaînes latérales : SCWRL, SCAP, NCN et FRMSCMFT a été effectuée sur une même base test (pour plus de détails, voir annexe C). Cette étude vise à répondre à trois questions : (i) Quel est le programme le plus performant ? (ii) Quelles sont les principales sources d'erreur ? (iii) Les prédictions sont-elles fiables lorsque la structure possède un squelette peptidique modélisé ?

Quel est le programme le plus performant ? Il est parfois délicat de comparer les différents programmes existants car les publications n'utilisent pas la même base de test ni les mêmes critères pour déterminer si une prédiction est correcte.

Quelles sont les principales sources d'erreur ? On souhaite savoir si des paramètres tels que l'accessibilité des chaînes latérales ou le type des résidus influencent les prédictions. Nous avons distingué trois types de résidus : les hydrophobes, les polaires et les chargés (cf. **table 15**) et trois niveaux d'accessibilité pour les chaînes latérales des résidus : $acc < 10\%$, $acc < 30\%$, et $acc \leq 100\%$ (ce qui englobe tous les résidus, aussi bien enfouis qu'exposés au solvant).

Abréviation	Description	Acides Aminés
H	Hydrophobes	W Y V I L M F
P	Polaires	C T S N Q
C	Chargés	D E R K H

table 15 : Les acides aminés sont répartis en trois catégories en fonction de leurs propriétés physico-chimiques. L'alanine, la glycine et la proline ne sont pas référencés dans la table car ils ne sont pas pris en compte lors de cette étude préliminaire.

Les prédictions sont-elles fiables lorsque la structure possède un squelette peptidique modélisé ? Les programmes d'optimisation du placement des chaînes latérales sur un squelette fixe sont fréquemment utilisés dans le but de repositionner au mieux les chaînes latérales sur un squelette modélisé. Une grande partie des travaux publiés mesure la fiabilité des méthodes en effectuant des prédictions en aveugle sur des structures de protéines expérimentales très précises. Il est donc important d'explorer et de quantifier la qualité des prédictions correctes lorsqu'on effectue la prédiction à partir d'un squelette non expérimental, issu par exemple d'une modélisation prédictive de la structure.

5.2.2 Base de référence.

La base de structures de référence est composée de 31 structures issues de la PROTEIN DATA BANK (Berman et al., 2000) déterminées par cristallographie, ayant une seule chaîne, une résolution $r \leq 1.8$ Å et une identité de séquence 2 à $2 \leq 50\%$ (voir **figure 59**).

Puisque nous souhaitons évaluer l'impact de la précision du squelette modélisé sur les capacités prédictives des différents programmes, nous avons besoin de travailler à la fois sur des structures dont le squelette est exact, mais aussi sur des structures sont le squelette dévie plus ou moins de sa position dans la structure expérimentale.

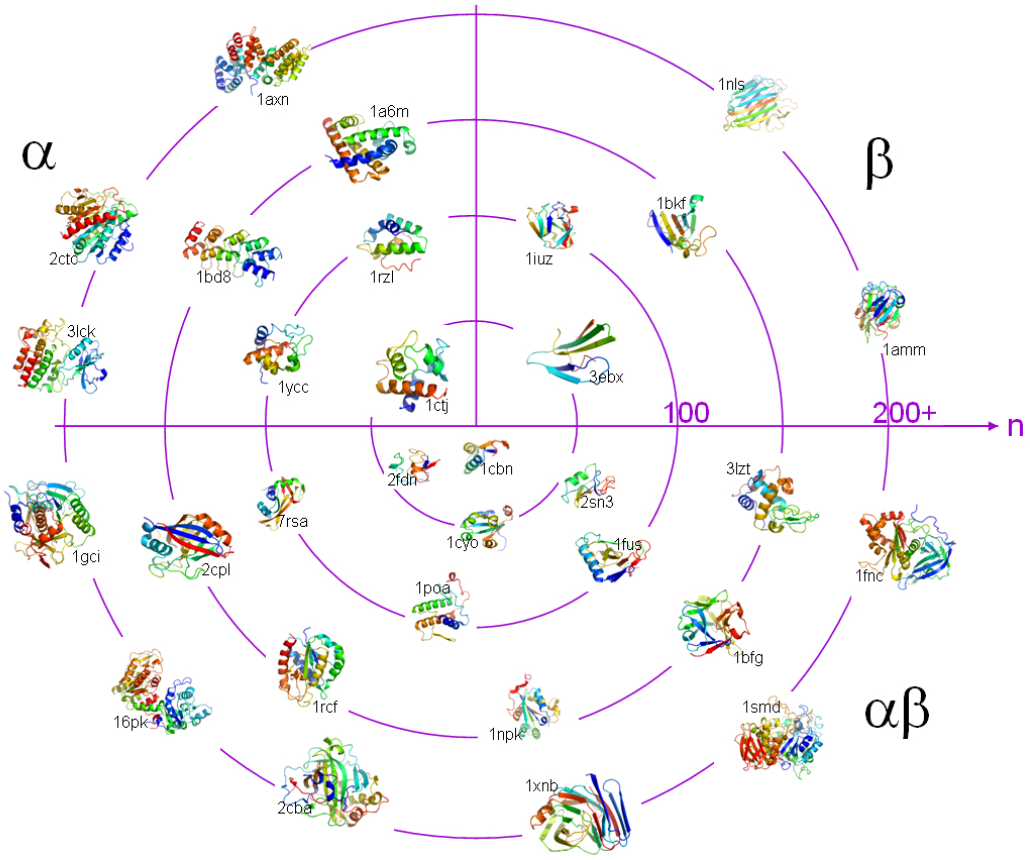


figure 59: Les 31 structures résolues par diffraction aux rayons X qui constituent notre base de référence. Les différentes structures sont placées sur des cercles concentriques en fonction de leur longueur en acides aminés et de leur composition en structure secondaire : le quadrant en haut à gauche contient toutes les structures uniquement en hélice α, le quadrant en haut à droite contient les structures en brins β, et le bas est constitué des structures contenant à la fois des hélices α et des feuillets β.

Quatre pools de structures ont été considérés :

- Le **pool1,** constitué des structures expérimentales (obtenues par cristallographie) élaguées de leurs chaînes latérales.
- Le **pool2,** constitué de ces mêmes structures expérimentales minimisées avant l'élagage des chaînes latérales. De cette façon, nous avons constaté que l'écart quadratique moyen (rmsd) entre les squelettes des structures du pool1 et celles du pool2 était d'environ 0.5 Å.
- Les pools 3 et 4 sont quant à eux constitués de modèles des structures de référence construits par homologie. Le **pool3** regroupe des modèles très proches de la structure expérimentale à laquelle ils se rapportent (rmsd entre 0 et 2 Å), minimisés puis élagués de leurs chaînes latérales. Pour le **pool4**, les modèles considérés sont moins proches (rmsd entre 2 et 4 Å).

Les détails de la construction de la base de structures de référence et des quatre pools de structures sont expliqués dans la partie Matériel et Méthodes (Annexe C).

5.2.3 Taux de prédictions correctes sur squelettes non modélisés (pools 1&2) en fonction de la tolérance angulaire.

Pour discriminer si les angles de torsion prédits pour les chaînes latérales sont corrects, une valeur seuil de tolérance angulaire ε doit être définie, telle qu'un angle soit considéré comme correct si et seulement si il dévie d'au plus ε degrés par rapport à sa valeur dans la structure de référence.

Dans un premier temps, l'impact de la tolérance angulaire sur le taux de prédiction des différents programmes a été étudié. Il est en effet important d'exclure un biais dû au choix de la tolérance angulaire dans notre comparaison des différents programmes. La **figure 60** présente le taux de prédictions correctes en fonction de la tolérance angulaire ε. Afin ne pas introduire de bruit dans cette analyse, seules les structures des pools 1 et 2 ont été utilisées (c'est-à-dire les structures dont le squelette ne résulte pas d'une modélisation par homologie).

Sur le graphe représentant le taux de prédictions correctes d'angles χ_1 en fonction de la tolérance angulaire, on observe quatre courbes quasi-identiques. Pour les angles χ_{12}, cad. χ_1 et χ_2 corrects simultanément, le programme FRMSCMFT est légèrement moins performant que les trois autres programmes dès que la tolérance angulaire dépasse 20°. Ce résultat peut être dû à un problème d'échantillonnage : au vue du nombre de rotamères et de sous rotamères pris en compte, l'espace des conformations des angles χ_1 et χ_2 à explorer est probablement trop important. Malgré cela, on note que les quatre courbes ont des allures très similaires.

En conclusion, on constate que le choix d'une valeur précise de tolérance angulaire ε a une incidence très faible sur l'évaluation comparée des différents programmes les uns relativement aux autres. En revanche, ce paramètre qui varie d'une publication à l'autre a une incidence directe sur les taux de prédictions correctes calculés. Dans la suite de l'étude comparative, la tolérance angulaire ε a été fixée à 20°.

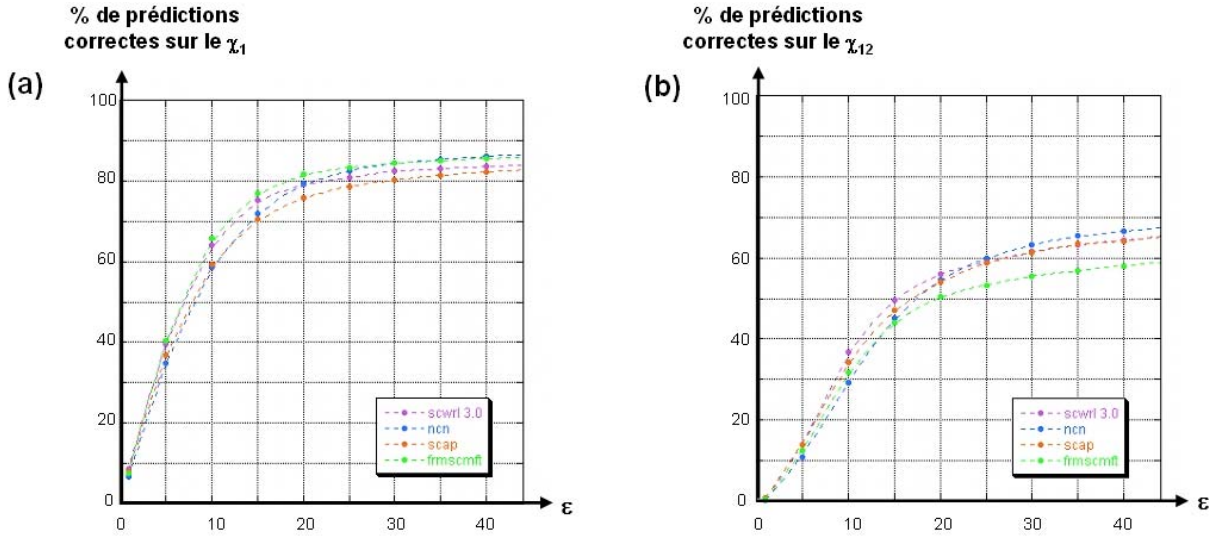


figure 60 : Influence de la tolérance angulaire sur le taux de prédictions correctes des programmes SCWRL (courbe violette), NCN (courbe bleue), FRMSCMFT (courbe verte) et SCAP (courbe orange). **(a)** La courbe représente le taux d'angles χ_1 correctement prédits. **(b)** La courbe représente le taux d'angles χ_{12} correctement prédits : c.a.d. χ_1 correct et χ_2 correct.

5.2.4 Taux de prédiction en fonction de l'erreur commise lors de la modélisation du squelette.

Afin de mesurer l'impact global de l'utilisation d'un squelette peptidique modélisé par homologie à la place d'un squelette peptidique expérimental, les angles prédits χ_1 et χ_{12} ont été comparés aux angles dièdres de la structure expérimentale sur les quatre pools de données. Les résultats obtenus sont présentés dans la **table 16**.

	angle χ_1				angles χ_{12}			
méthode	**pool 1**	**pool 2**	**pool 3**	**pool 4**	**pool 1**	**pool 2**	**pool 3**	**pool 4**
SCWRL 3.0	0.79	0.79	0.67	0.62	0.56	0.54	0.44	0.41
NCN	0.79	0.77	0.61	0.58	0.55	0.54	0.39	0.36
SCAP	0.76	0.75	0.61	0.56	0.54	0.54	0.39	0.35
FRMSCMFT	0.82	0.80	0.68	0.63	0.50	0.51	0.40	0.39

table 16 : Analyse des taux de prédictions correctes des programmes SCWRL, FRMSCMFT, SCAP et NCN en fonction du pool de structures pour les angles χ_1 (à gauche) et χ_{12} (à droite). Le **pool1** rassemble des structures expérimentales élaguées de leurs chaînes latérales. Le **pool2** rassemble les mêmes structures expérimentales minimisées (rmsd ≈ 0.5 Å). Le **pool3** est l'ensemble des modèles très fiables construits par homologie à partir des mêmes structures (0Å ≤ rmsd < 2Å). Le **pool4** regroupe les modèles moins fiables (2Å ≤ rmsd < 4Å).

Comme on pouvait le supposer, les taux de prédiction se dégradent lorsqu'on s'éloigne du squelette peptidique expérimental. Ceci est valable quelle que soit la méthode de prédiction utilisée. Sur les squelettes expérimentaux, le taux de χ_1 corrects se situe autour de 80%, et le taux de χ_{12} autour de 55%. On ne constate pas de différences sensibles entre les différentes méthodes ; si ce n'est le problème de FRMSCMFT pour la prédiction des angles χ_{12} déjà évoqué. Les performances atteintes pour le pool 2 sont assez proches de celles observées sur les squelettes expérimentaux. Ainsi, on peut constater que les faibles perturbations du positionnement du squelette n'influent pas significativement sur les prédictions. Sur les squelettes modélisés (pools 3 et 4), les taux de prédiction chutent sensiblement : -20%du pouvoir prédictif pour les angles χ_1, et environ -30% pour les angles χ_{12} , quelle que soit la méthode utilisée. Ces résultats mettent en évidence les limitations actuelles des méthodes de prédiction de la conformation des chaînes latérales sur des squelettes peptidiques modélisés par homologie : 60% des angles χ_1 et 40% des angles χ_{12} sont corrects.

Néanmoins, la similitude structurale est distribuée de façon inhomogène au sein des modèles construits par homologie : il est probable que seules certaines régions soient bien modélisées. Nous avons donc étudié le taux de prédictions correctes des angles χ_1 et χ_{12} en fonction de la distance Cα-Cα entre la structure expérimentale et le modèle lorsqu'ils sont superposés. Ainsi, on passe d'une mesure d'erreur globale (s'appliquant à une structure entière), à une mesure d'erreur locale (s'appliquant à 1 résidu d'une structure donnée). Les résultats sont présentés sur la **figure 61**.

Aucun des quatre programmes testés ne se démarque des autres. Les taux de prédiction du χ_1 ou du χ_{12}, ont la même allure pour les quatre courbes. Lorsque la distance Cα-Cα varie de 0.0 à 0.5 Å, la perte de pouvoir prédictif est significative : c'est en effet sur cette portion de la courbe que la pente est la plus forte. Cette propriété est retrouvée aussi bien pour les angles χ_1 que pour les angles χ_{12}, et quelle que soit la méthode de prédiction de la conformation des chaînes latérales utilisée (**figure 61**-B-D). Le taux de prédiction continue de décroître de façon assez régulière entre 0.5 et 2.0 Å. Enfin, alors qu'entre 0.0 et 2.0 Å la courbe était décroissante de façon quasi-monotone (**figure 61**-B-D), on constate que lorsque la distance Cα-Cα dépasse 2.0 Å, celle-ci devient assez irrégulière (**figure 61**-A-C).

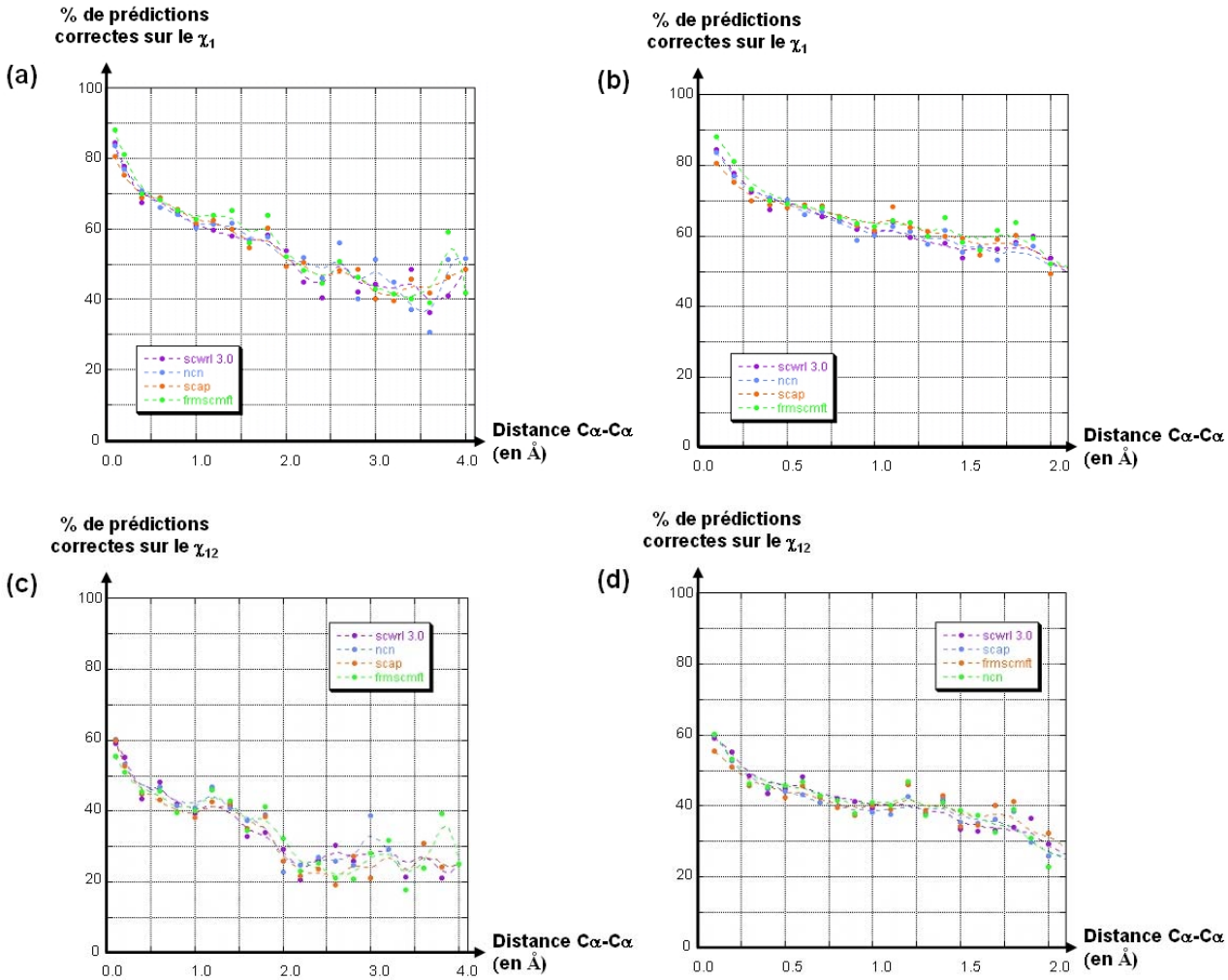


figure 61 : Graphes représentant le taux des prédictions correctes des angles dièdres décrivant les chaînes latérales en fonction des distances Cα-Cα. Chaque structure est superposée sur la structure expérimentale correspondante. Ensuite, on calcule, pour chaque résidu, la distance Cα-Cα. **(a)** Taux d'angles χ_1 corrects en fonction de la distance Cα-Cα, sur l'intervalle [0-4] Angstroms. **(b)** Idem, agrandissement de l'intervalle [0-2] Angstroms. **(c)** Taux d'angles χ_{12} corrects en fonction de la distance Cα-Cα, sur l'intervalle [0-4] Angstroms. **(d)** Idem, agrandissement de l'intervalle [0-2] Angstroms.

5.2.5 Taux de prédictions correctes en fonction du type de résidus et de l'accessibilité.

Nous étudions maintenant l'impact du type de résidu (hydrophobe, polaire, chargé) et de l'accessibilité des chaînes latérales sur le taux de prédiction des différents programmes.

La **table 17** synthétise les résultats obtenus sur les pools 1 et 2. En prenant en compte les résultats sur l'ensemble des résidus (cf. colonnes 100%), on constate que la prédiction de la conformation des chaînes latérales des résidus hydrophobes est plus efficace que celle des résidus chargés ou polaires. Par exemple, pour le programme NCN, le taux d'angles χ_1 correctement prédits est de 88% pour les hydrophobes, 75% pour les chargés et 73% pour les polaires. Pour les taux de prédiction du χ_{12}, le pourcentage d'angles correctement prédits est de 65% pour les résidus hydrophobes, de 44% pour les chargés et de 47% pour les polaires. De la même manière, quel que soit le programme utilisé, la prédiction de la conformation des chaînes latérales des résidus hydrophobes est meilleure que celle des autres résidus. De plus, la différence de taux de prédiction, déjà importante au niveau de l'angle χ_1, est accentuée significativement lorsqu'on considère les taux de prédiction des angles χ_{12}.

L'une des hypothèses pouvant expliquer cette différence entre hydrophobes d'un côté et polaires et chargés de l'autre est que la prédiction de la conformation des chaînes latérales est plus facile dans le cœur des protéines car le volume y est contraint et que les hydrophobes sont les acides aminés les plus abondants dans le cœur des protéines.

Pourtant, en restreignant notre analyse aux résidus appartenant au cœur des protéines (cf. colonne 10%), on constate que les taux de prédiction des résidus hydrophobes sont toujours meilleurs que ceux des résidus polaires et chargés. Ceci est vrai au niveau du taux de prédictions correctes des angles χ_1 (**table 17**, colonnes en vert), et au niveau des angles χ_{12} (**table 17**, colonnes en bleu). A titre d'exemple, les performances du programme SCWRL sont de 91% de χ_1 corrects pour les résidus hydrophobes enfouis, contre 81% pour les chargés enfouis et 84% pour les polaires enfouis ; tandis qu'au niveau du taux de prédictions correctes des angles χ_{12}, on passe de 70% pour les hydrophobes enfouis à 53% pour les chargés enfouis et 60% pour les polaires enfouis. Sur les pools 3 et 4, correspondant aux structures modélisées, les résultats suivent la même tendance (**table 18**).

Nos analyses suggèrent donc que le positionnement des acides aminés hydrophobes serait plus efficace que celui des polaires et chargés pour des raisons en partie indépendantes de l'accessibilité.

χ_1	H			P			C		
méthode	10%	30%	100%	10%	30%	100%	10%	30%	100%
SCWRL 3.0	0.91	0.89	0.88	0.84	0.78	0.73	0.81	0.77	0.76
NCN	0.92	0.90	0.88	0.87	0.80	0.73	0.82	0.77	0.75
SCAP	0.93	0.91	0.89	0.84	0.74	0.68	0.80	0.73	0.72
FRMSCMFT	0.94	0.91	0.90	0.84	0.80	0.74	0.85	0.80	0.78

χ_{12}	H			P			C		
méthode	10%	30%	100%	10%	30%	100%	10%	30%	100%
SCWRL 3.0	0.70	0.65	0.64	0.60	0.56	0.50	0.53	0.46	0.41
NCN	0.71	0.67	0.65	0.60	0.55	0.47	0.51	0.48	0.44
SCAP	0.80	0.77	0.75	0.52	0.48	0.44	0.56	0.38	0.37
FRMSCMFT	0.70	0.66	0.65	0.31	0.38	0.37	0.46	0.38	0.35

table 17 : Analyse des taux de prédictions correctes des programmes SCWRL, NCN, SCAP et FRMSCMFT sur des structures dont le squelette n'est pas modélisé (pools 1 et 2), en fonction de la nature des résidus et de l'accessibilité des chaînes latérales. La table du haut donne les résultats pour le taux de prédiction correcte des angles χ_1. La table du bas donne le même taux pour les angles χ_{12}, c'est-à-dire χ_1 correct et χ_2 correct. Les types de résidus sont hydrophobes (H), polaires (P) et chargés (C). L'accessibilité des chaînes latérales des résidus peut être soit inférieure ou égale à 10%, correspondant au cœur enfoui des protéines, soit inférieure ou égale à 30%, soit inférieure ou égale à 100%, englobant tous les résidus quelle que soit leur accessibilité.

χ_1	H			P			C		
méthode	10%	30%	100%	10%	30%	100%	10%	30%	100%
SCWRL 3.0	0.75	0.74	0.74	0.63	0.62	0.61	0.64	0.58	0.57
NCN	0.78	0.74	0.74	0.60	0.60	0.59	0.63	0.58	0.57
SCAP	0.77	0.75	0.75	0.63	0.61	0.61	0.65	0.60	0.60
FRMSCMFT	0.79	0.76	0.76	0.61	0.60	0.59	0.64	0.60	0.60

χ_{12}	H			P			C		
méthode	10%	30%	100%	10%	30%	100%	10%	30%	100%
SCWRL 3.0	0.55	0.54	0.53	0.38	0.33	0.33	0.41	0.35	0.35
NCN	0.56	0.53	0.53	0.40	0.33	0.33	0.40	0.35	0.35
SCAP	0.53	0.52	0.51	0.36	0.30	0.30	0.38	0.32	0.31
FRMSCMFT	0.54	0.52	0.52	0.40	0.31	0.31	0.37	0.32	0.32

table 18 : Analyse des taux de prédictions correctes des programmes SCWRL, NCN, SCAP et FRMSCMFT sur des structures dont le squelette est modélisé (pools 3 et 4), en fonction de la nature des résidus et de l'accessibilité des chaînes latérales. Légende identique à **table 17**.

Nous suggérons l'hypothèse selon laquelle le positionnement des résidus polaires et chargés enfouis est plus délicat en raison de l'inter-dépendance du positionnement local de chacun de ces résidus. En effet, dans le cœur des protéines, les résidus polaires et chargés doivent trouver des partenaires afin de former des réseaux de liaisons hydrogènes complexes car il est très défavorable d'enfouir un atome polaire sans que celui-ci ne forme une liaison hydrogène. Ainsi, pour prédire correctement le positionnement des acides aminés polaires et chargés du cœur des protéines, une erreur sur le positionnement d'un seul résidu remet en cause tout le réseau de liaisons hydrogènes auquel il participe, selon le principe du « tout ou rien ». A l'inverse, dans le cas des amas hydrophobes, une erreur sur le positionnement d'un des résidus a moins d'impact car il peut être facilement compensé localement. Au chapitre 6,

nous retrouverons cette difficulté prédictive associée aux chaînes latérales polaires et chargées avec l'exemple d'une arginine enfouie dans l'interface entre un domaine FHA et son ligand.

5.2.6 Conclusion.

Au cours de cette étude préliminaire, nous avons pu mettre en évidence que les quatre méthodes comparées possédaient des taux de prédictions correctes proches et partageaient les mêmes limitations. Ce résultat contraste avec les conclusions qui pouvaient être établies à la lecture des publications de chacun de ces programmes et souligne l'importance du test préliminaire effectué.

Cette étude visait à sélectionner un programme de prédiction de la conformation des chaînes latérales dans le but de l'utiliser pour modéliser les interfaces des différents mutants qui seront évalué par la fonction d'énergie FOLDEF. Au vu des résultats obtenus, aucune méthode ne semble significativement plus efficace que les autres. Nous avons donc choisi le programme en nous basant sur des critères de (i) granulosité de la bibliothèque de rotamères et (ii) rapidité d'exécution. Le programme SCAP réalisait le meilleur compromis entre ces deux aspects et a donc été sélectionné (**figure 62**) (voir également **Article 3**).

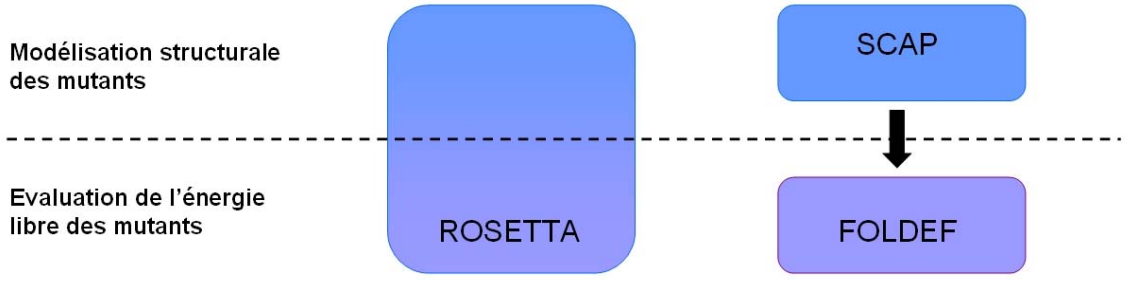


figure 62 : Deux méthodes de criblage *in silico*. La première méthode utilise la fonction de score ROSETTA, et possède son propre moteur de recherche pour la modélisation des mutants. La seconde approche utilise la fonction de score FOLDEF et est couplée au programme SCAP permettant l'exploration conformationnelle pour la modélisation des mutants.

5.3 Criblage *in silico* des domaines FHA et des tandems de domaines BRCT.

5.3.1 Méthdologies testées.

Cinq structures constituent notre base de test : le domaine FHA N-terminal de Rad53, le domaine FHA de Chk2, le domaine FHA de Pnk, et les tandems de domaines BRCT des protéines Brca1 et Mdc1.

Quatre stratégies ont été explorées pour le criblage *in silico* : (1) l'utilisation de ROSETTA seul, (2) le couplage SCAP + FOLDEF, (3) l'évaluation de l'énergie par FOLDEF sur les structures modélisées par ROSETTA, (4) utilisation de la moyenne des énergies FOLDEF et ROSETTA sur les structures modélisées par ROSETTA. La valeur du ΔΔG est évaluée en prenant comme référence la position en alanine.

Dans la section 5.3.2, les méthodologies sont appliquées sur le domaine FHA N-terminal de Rad53 à titre d'illustration. Concernant les autres domaines, les résultats complets sont présentés en annexe Annexe D et synthétisés dans la partie 5.3.3.

5.3.2 Criblage *in silico* du domaine FHA N-terminal de Rad53 sur la position pT+3.

Le criblage *in silico* du domaine FHA N-terminal de Rad53 en position pT+3 par ROSETTA sélectionne le motif recherché pTxxD ($\Delta\Delta G$ = -2,5 kCal.mol^{-1}) mais aussi d'autres motifs comme pTxxF ($\Delta\Delta G$ = -2,0 kCal.mol^{-1}) ou pTxxW ($\Delta\Delta G$ = -1.8 kCal.mol^{-1}). Dans l'ensemble et à l'exception du motif pTxxD, la présence de résidus hydrophobes en position pT+3 semble favorable. Cette propriété n'est pas vérifiée par la seconde méthode de criblage basée sur la modélisation par SCAP et l'évaluation par FOLDEF. Les résidus hydrophobes sont prédits comme défavorables tandis que les résidus chargés négativement sont sélectionnés avec une nette préférence pour l'aspartate ($\Delta\Delta G$ = -2,5 kCal.mol^{-1}) plutôt que pour le glutamate ($\Delta\Delta G$ = -1,6 kCal.mol^{-1}).

Nous nous sommes interrogés sur cette différence d'appréciation des résidus hydrophobes en position pT+3 par les deux approches. Celle-ci peut avoir comme origine soit (i) le fait que de meilleures solutions ont été générées lors de la phase d'exploration conformationnelle de ROSETTA ; soit (ii) une évaluation différente par ROSETTA et FOLDEF de structures proches.

Afin de répondre à cette question, les structures modélisées par ROSETTA ont été évaluées par FOLDEF (**figure 63-c**). La comparaison des **figure 63-a** et **figure 63-c** met en avant la différence entre ROSETTA et FOLDEF sur des structures identiques. Parmi les écarts d'évaluation les plus importants se trouve la structure associée au motif proposant un tryptophane en position pT+3 : la structure de ce mutant pTxxW est prédite comme significativement favorable par ROSETTA ($\Delta\Delta G$ = -1,8 kCal.mol^{-1}) tandis que FOLDEF ne sélectionne pas cette structure ($\Delta\Delta G$ = -0.05 kCal.mol^{-1}).

La comparaison des **figure 63-b** et **figure 63-c** souligne la différence de qualité de modélisation entre SCAP et ROSETTA. Le motif le plus affin est pTxxD dans les deux cas, et l'énergie libre associée aux deux complexes est proche (-3.0 kCal.mol^{-1} < $\Delta\Delta G$ < -2.0 kCal.mol^{-1} dans les deux cas). Si on considère les acides aminés prédits comme défavorables par rapport à l'alanine ($\Delta\Delta G$ positif) par l'approche SCAP+FOLDEF, on remarque que ces mêmes acides aminés se retrouvent plutôt bien évalués par l'approche ROSETTA+FOLDEF. Cette tendance suggère que les interfaces modélisées par ROSETTA présentent des caractéristiques beaucoup plus favorables que celles modélisées par SCAP. Ceci met clairement en évidence un défaut d'exploration conformationnelle lorsqu'on utilise SCAP.

L'ensemble de ces résultats suggère que la conformation des chaînes latérales dans les mutants modélisés par ROSETTA est de meilleure qualité que ceux générés par SCAP, bien que la fonction d'énergie FOLDEF soit plus performante dans son évaluation comparée des différents mutants.

Pour conclure, une moyenne des énergies estimées par les deux fonctions d'évaluation sur les structures générées par ROSETTA est présentée sur la **figure 63-d**. D'après celle-ci, les motifs les plus affins sont pTxxD ($\Delta\Delta G$ = -2,65 kCal.mol^{-1}), pTxxF ($\Delta\Delta G$ = -1,91 kCal.mol^{-1}) et dans une moindre mesure pTxxM ($\Delta\Delta G$ = -1,62 kCal.mol^{-1}). Ces résultats *in silico* sont en accord avec les données expérimentales *in vitro* puisque seul le motif pTxxD a été identifié comme le plus affin.

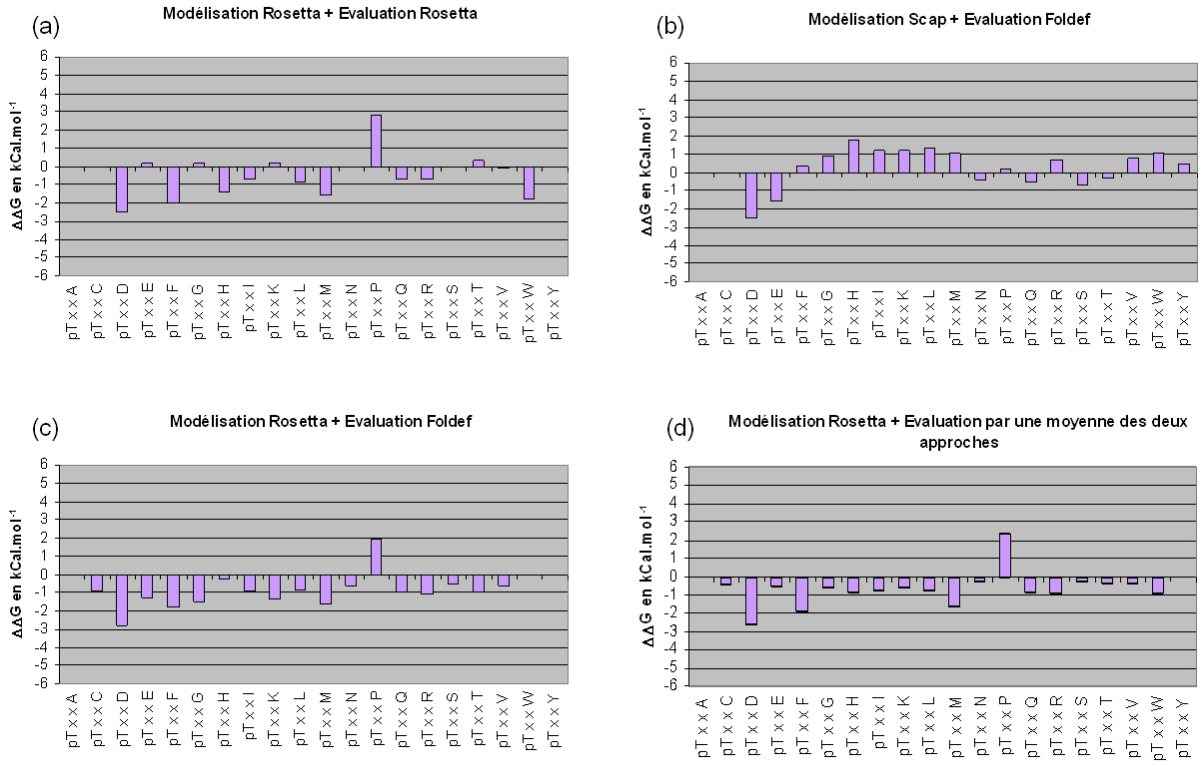


figure 63 : Criblage in silico du domaine FHA N-terminal de la protéine de levure Rad53 sur la position pT+3. Le motif le plus affin pour ce domaine FHA est pTxxD. Pour chaque méthode, l'abscisse représente les motifs testés et l'ordonnée la différence d'énergie libre estimée entre le motif et le motif pTxxA. **(a)** Modélisation Rosetta, Evaluation Rosetta. **(b)** Modélisation Scap, Evaluation Foldef. **(c)** Modélisation Rosetta, Evaluation Foldef. **(d)** Moyenne des énergies calculées en (a) et en (c).

5.3.3 Criblage *in silico* des domaines FHA et des tandems de domaines BRCT restreint à la position spécifiquement reconnue.

Pour les cinq structures de complexes criblées expérimentalement, le motif spécifiquement reconnu déterminé expérimentalement est toujours détecté par au moins l'une des quatre approches (**table 19**).

Les résultats constatés sont assez paradoxaux. D'un côté, le motif spécifiquement reconnu est quasi-systématiquement placé dans le peloton de tête des prédictions, et souvent à la première place (**table 19-colonne 3)**. Par exemple, dans le cas du domaine FHA N-terminal de Rad53 détaillé au paragraphe précédent, les quatre approches coïncident sur le fait que le motif pTxxD est le plus affin. Ce résultat très positif est cependant contrasté par un défaut de sélectivité. En effet, en prenant en compte la marge d'erreur des fonctions d'évaluation (environ 0,80kCal.mol^{-1}), on constate qu'un certain nombre de faux-positifs peuvent être

décelés : l'écart entre l'énergie libre estimée pour ces motifs et l'énergie libre estimée pour le motif reconnu expérimentalement est inférieure à 0,80kCal.mol^{-1}. Ces résultats suggèrent que les capacités discriminantes des méthodes employées doivent être améliorées.

Structure	Motif le plus affin (détection expérimentale)	Approches qui détectent ce motif comme le plus affin	Approches qui détectent ce motif comme le plus affin avec une marge de 0,80 kCal.mol^{-1}
FHA N-terminal de Rad53	pTxxD	(1) (2) (3) (4)	(2) (3) (4)
FHA de Chk2	pTxx[I/L]	(1) (3) (4)	
FHA de Pnk	DxxpT	(2) (3) (4)	
Tandem BRCT de Brca1	pSxx[Y/F]	(2) (3) (4)	(2) (3) (4)
Tandem BRCT de Mdc1	pSxx[Y/F]	(2) (3)	(2) (3)

table 19 : Table récapitulant les résultats obtenus par les 4 approches testées (résultats détaillés en annexe C). Méthodes (1) modélisation par ROSETTA + évaluation par ROSETTA ; (2) modélisation par SCAP + évaluation par FOLDEF ; (3) modélisation par ROSETTA + évaluation par FOLDEF ; (4) modélisation par ROSETTA + évaluation par une moyenne des énergies prédites par ROSETTA et FOLDEF.

On constate que l'approche utilisant ROSETTA pour la modélisation des mutants et une moyenne de ROSETTA et FOLDEF pour leur évaluation est efficace bien qu'elle manque de sélectivité. Sur les cinq exemples testés avec cette approche, le résidu spécifiquement reconnu est classé comme le plus affin dans quatre cas : FHA1 de Rad53, FHA de Chk2, FHA de Pnk, tandem BRCT de Brca1, bien que les marges d'erreur des méthodes d'évaluation rendent possible la détection de faux-positifs. Pour le dernier exemple traité, celui du tandem BRCT de la protéine Mdc1, les motifs pSxxY et pSxxF reconnus spécifiquement sont placés respectivement à la deuxième et à la troisième place, derrière le motif pTxxW. Le motif placé est première position n'est donc pas celui reconnu expérimentalement, mais il partage néanmoins le caractère aromatique des résidus spécifiquement reconnus.

5.3.4 Criblage *in silico* des domaines FHA et des tandems de domaines BRCT sur toute la longueur des peptides.

Dans ce paragraphe, les cinq exemples précédents sont criblés sur toute la longueur du peptide et plus seulement sur la position dont les expériences ont montré qu'elle était responsable de la spécificité. Nous avons sélectionné une des stratégies les plus performantes identifiée dans la section précédente et reposant sur le calcul de la moyenne des énergies

FOLDEF et ROSETTA sur les structures modélisées par ROSETTA. Le but de ce criblage sur toute la longueur du peptide est donc avant tout de mettre en évidence la capacité des méthodes de *design* à identifier les positions les plus sélectives. Notons que la procédure de criblage ne sera pas appliquée aux positions du peptide exposées au solvant et n'ayant pas de contacts avec les domaines criblés.

Structure	Position déterminée comme spécifique expérimentalement	Positions prédites comme spécifiques
FHA N-terminal de Rad53	pT+3	pT-4 , pT+3
FHA de Chk2	pT+3	pT-3 , pT-2 , pT+1, pT+2, pT+3
FHA de Pnk	pT-3	pT-4 , pT-2
Tandem de BRCT de Brca1	pS+3	pS+1 , pS+2 , pS+3
Tandem de BRCT de Mdc1	pS+3	pS+2 , pS+3

table 20 : Table récapitulant les position spécifique calculées par criblage in silico pour les cinq structures de départ.

Les résultats complets de ce crible sont présentés en détails au sein de l'annexe D. La **table 20** synthétise les résultats obtenus. Pour chaque structure, les positions présentant un profil de « position spécifique » ont été recensées. Dans tous les exemples étudiés, il existe une ou plusieurs positions pour lesquelles certains acides aminés sont prédits comme plus favorables. Or pour ces positions, les données expérimentales n'indiquaient aucune préférence particulière. Cette situation est problématique car dans un test en aveugle elle aurait conduit à définir un motif consensus erroné. Cependant, rappellons tout de même qu'au vu des résultats précédents, ce motif consensus aurait au moins contenu l'acide aminé effectivement responsable de la spécificité.

5.4 Conclusions et Perspectives.

L'évaluation des performances des méthodes de *design* pour prédire les motifs spécifiquement reconnus apparaît donc tout à fait favorable dans des cas de figure où le squelette peptidique peut légitimement être maintenu rigide. Cette situation englobe un grand nombre de familles de PRMs dans lesquelles le site de liaison est assuré par des éléments de structures secondaires *a priori* rigides. Citons par exemple le cas des domaines PDZ.

Les PRMs évalués ici, domaines FHA et BRCT, sont particulièrement intéressants puisqu'ils reconnaissent une grande variété de fonctions chimiques : polaires, aromatiques, chargés, ou hydrophobes. Les résultats obtenus dans ce chapitre présentent donc un intérêt général pour l'étude de l'ensemble des PRMs.

Quatre stratégies ont été explorées pour le criblage *in silico* : (1) ROSETTA seul, (2) SCAP + FOLDEX, (3) évaluation FOLDEF sur les structures modélisées par ROSETTA, (4) moyenne des énergies FOLDEF et ROSETTA les structures modélisées par ROSETTA. Les résultats obtenus suggèrent que l'approche utilisant ROSETTA pour la modélisation des mutants et une moyenne des fonctions d'énergie de ROSETTA et FOLDEF apparaît comme la plus spécifique. Pour quatre des cinq exemples testés avec cette approche, l'acide aminé spécifiquement reconnu a été correctement prédit. Notons toutefois que les marges d'erreur des méthodes d'évaluation rendent possible la détection de faux-positifs.

Les performances des deux programmes de *design* de protéines utilisés, ROSETTA et FOLDEF, apparaissent relativement similaires. Les résultats suggèrent ici que la fonction d'énergie FOLDEF est plus fiable. Néanmoins, dans une étude comparée de plusieurs méthodes sur d'autres cibles (**Article 3**), nos conclusions étaient plus favorables au programme ROSETTA.

Il est important de noter que ces deux méthodes ont été développées pour prédire correctement la contribution énergétique d'interactions observées expérimentalement dans les structures. Dans les deux cas, l'apprentissage a été effectué sur des bases de données de mutants pour lesquels on disposait de données thremodynamiques. Il n'a donc pas explicitement intégré de contraintes pour distinguer une interaction exacte d'une interaction fausse. Cette notion peut expliquer que les deux méthodes présentent toutes les deux une bonne sensibilité (le bon motif est toujours parmi les meilleurs) mais une spécificité limitée. L'intérêt d'utiliser une moyenne des deux fontions d'énergie peut s'expliquer par le fait que les faux-positifs détectés par chaque approche ne sont pas les mêmes. Cette solution est n'est toutefois pas satisfaisante car lorsqu'on considère le criblage en dehors des positions reconnues spécifiquement (5.3.3), les autres positions présentent des préférences assez fortes en contradiction avec les résultats expérimentaux.

Cette étude, outre son intérêt pour la prédiction des spécificités d'interactions, fournit des pistes originales pour améliorer les performances des fonctions d'énergie telles que FOLDEF et ROSETTA. La limitation majeure de ces fonctions d'énergie est leur manque de sélectivité. L'introduction de faux positifs dans la base d'apprentissage doit permettre d'atténuer ce défaut de sélectivité. L'ensemble des structures générées au cours de ces cribles pourrait être utilisé dans cette optique.

Chapitre 6 : Prédiction des spécificités de reconnaissance des PRMs : prise en compte des mouvements du squelette.

Dans ce dernier chapitre, nous étudions la possibilité d'introduire des mouvements au niveau du squelette peptidique lors du processus de criblage in silico. L'objectif à long terme est de parvenir à cribler un PRM à partir d'une structure plus ou moins éloignée de sa conformation complexée.

La simulation de dynamique moléculaire a été utilisée afin d'explorer l'espace des conformations accessibles à l'interface entre le PRM et son ligand peptidique. A travers cette étude, nous avons identifié les difficultés qui limitent actuellement le développement d'outils de prédiction fiables. Ce constat a stimulé la conception de nouvelles solutions algorithmiques qui lèvent une partie des verrous méthodologiques rencontrés.

6.1

Le problème de la flexibilité de l'interface protéine-peptide.

6.1.1 Introduction.

Dans le processus de liaison entre un PRM et un peptide, la flexibilité des régions variables séparant les éléments de structure secondaire du PRM peut jouer un rôle important. C'est notamment le cas de certains domaines WW ou FHA (Ding et al., 2005; Peng et al., 2007). Dans des cas spécifiques, la dynamique de ces régions non structurées a été étudiée par Résonance Magnétique Nucléaire. Les résultats des deux études concordent et indiquent que certaines de ces régions, flexibles lorsque le domaine n'est pas complexé à son peptide, se rigidifient au contact de celui-ci. Dans le cas du domaine WW de la protéine humaine Pin1, c'est le cas de la boucle de reconnaissance ^{16}SRSSGR^{21}, dont l'arginine 17 est en contact direct avec le résidu pT/pS reconnu. De plus, le raccourcissement de cette boucle du domaine affaiblit l'affinité pour son substrat (Peng et al., 2007). Dans le cas du domaine FHA de la protéine d'*Arabidopsis thaliana* KAPP, une étude réalisée par l'équipe de Steven Van Doren (*University of Columbia*, Missouri, USA) a mis en évidence un certains nombre de résidus dont les propriétés dynamiques varient entre forme complexée et non complexée (Ding et al., 2005). J'ai projeté ces résultats sur l'alignement de séquence des domaines FHA (figure 64) ainsi que sur la structure du domaine FHA de KAPP (figure 64). On constate que les résidus proches de la position pT+3, responsable de la spécificité de reconnaissance, sont rigidifiés par l'interaction FHA/fragment protéique.

L'ensemble de ces résultats suggère que dans le cadre des domaines WW et des domaines FHA, il existe un lien fort entre (i) la dynamique de certaines régions non structurées situées à l'interface avec le fragment protéique reconnu et (ii) les mécanismes de la reconnaissance spécifique de ces fragments protéiques. Pour le criblage *in silico,* la flexibilité du squelette peptidique apparaît donc comme un paramètre important à prendre en compte. De plus, dans l'optique d'appliquer la stratégie de criblage *in silico* à des structures issues de modélisation comparative, il est crucial d'intégrer des mouvements plus conséquents que ceux des seules chaînes latérales tels qu'au chapitre précédent. Pour atteindre cet objectif, nous avons exploré l'intérêt de la simulation de dynamique moléculaire pour effectuer l'exploration conformationnelle du squelette lors du criblage *in silico*.

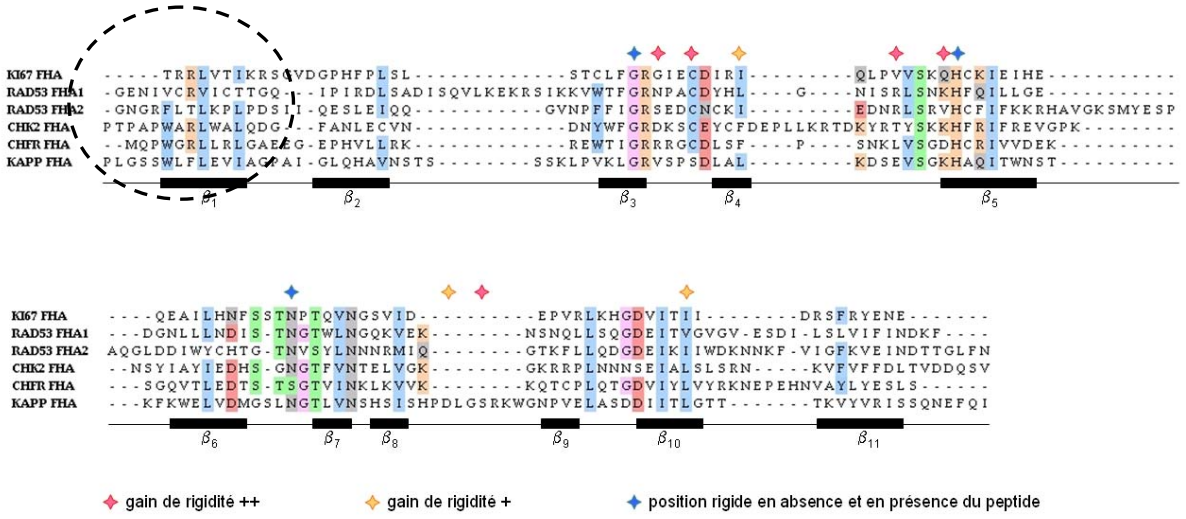


figure 64 : Alignement multiple des domaines FHA de structures connues sur lequel sont représentés les résidus dont les propriétés dynamiques varient en présence et en absence de peptide. Les losanges roses et jaunes indiquent les résidus qui se rigidifient respectivement beaucoup et un peu en présence du peptide. Les résultats sont projetés sur la structure (à droite) en utilisant le même code de couleurs.

Dans un premier temps, nous avons sélectionné un modèle d'étude PRM/ligand représentatif des difficultés d'exploration conformationnelle, avec par exemple un réseau complexe de liaisons hydrogènes stabilisant l'interface PRM/ligand. Le complexe formé par le domaine FHA1 de Rad53 et son peptide de plus forte affinité pTxxD est stabilisé par des liaisons hydrogènes inter- et intra-moléculaires. Certaines de ces liaisons, qui maintiennent le positionnement de la thréonine phosphorylée et du squelette du peptide, sont conservées dans tous les exemples connus de complexes FHA/peptide. Par contre, dans la poche de reconnaissance entourant l'aspartate en position pT+3, un réseau complexe de ponts salins et de liaisons hydrogènes spécifique au complexe FHA1 de Rad53 / peptide pTxxD est

observé. La présence de ce réseau de liaisons hydrogènes rend délicate l'exploration conformationnelle de cette interface. C'est la raison pour laquelle nous avons choisi ce complexe comme système d'étude. Les aspects structuraux cette interaction seront détaillées dans le paragraphe suivant (6.1.2).

Dans l'optique d'un criblage *in silico,* le problème du temps de calcul s'est avéré délicat à aborder et plusieurs pistes ont été abordées. Les résultats les plus intéressants ont été obtenus en guidant la dynamique par un ensemble de contraintes ambiguës qui incite à la formation de réseaux de liaisons hydrogènes et de larges amas hydrophobes. L'algorithme servant à définir ces contraintes ambiguës sera présenté dans le paragraphe 6.3.

6.1.2 Aspects structuraux de l'interaction entre le domaine FHA1 de Rad53 et un peptide pTxxD.

Le domaine FHA1 de la protéine Rad53 de *Saccharomyces cerevisiae* reconnaît préférentiellement les fragments protéiques pTxxD (Kd = 330 nM). La structure de ce complexe a été résolue (Durocher et al., 2000) : on constate que l'aspartate en position pT+3 qui est responsable en grande partie de la spécificité de reconnaissance est au cœur d'un réseau de liaisons hydrogènes particulièrement complexe.

La **figure 65**-A présente l'interface entre le domaine FHA1 de Rad53 et le peptide pTxxD. A titre de comparaison, la **figure 65**-B présente la même région du domaine FHA1 lorsque celui-ci n'est pas complexé. L'aspartate 8 du peptide spécifiquement reconnue forme un pont salin avec l'arginine 83 du domaine FHA1. Pour que ce pont salin soit formé, il faut que l'arginine 83 adopte une conformation où sa chaîne latérale pointe vers le cœur du domaine, alors qu'elle est exposée au solvant lorsque le domaine n'est pas complexé. Cette désolvatation de l'arginine 83 est compensée par la formation de plusieurs liaisons hydrogènes : deux liaisons avec l'aspartate du peptide, une liaison avec l'oxygène de la chaîne latérale de l'aspartate 139, et deux liaisons hydrogènes avec les oxygènes du squelette peptidique de la glycine 133 et de la valine 134. Ainsi, tous les atomes donneurs de l'arginine 83 ont trouvé un accepteur avec lequel former une liaison hydrogène, ce qui rend cette conformation énergétiquement favorable. La **figure 65**-C schématise plus particulièrement les liaisons hydrogènes mettant en jeu l'arginine 83.

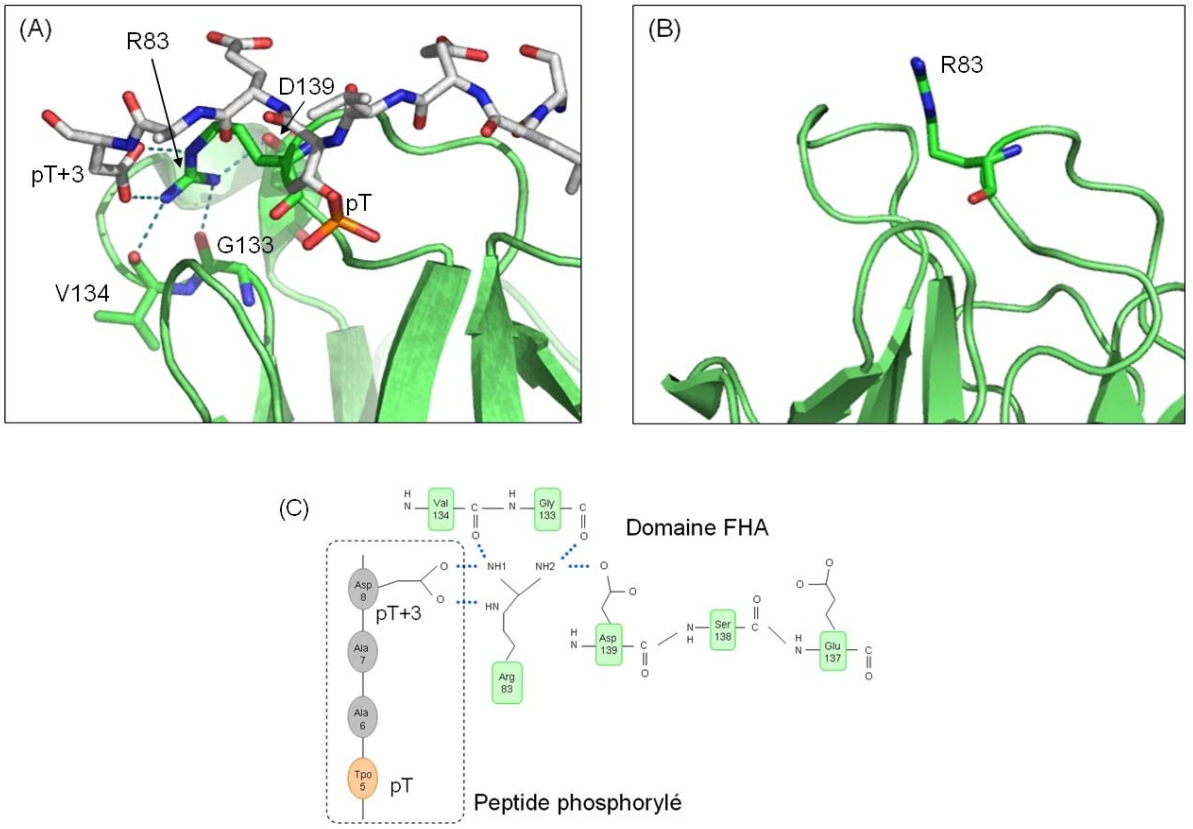


figure 65 : Organisation de la poche de reconnaissance entre l'aspartate du peptide et le domaine FHA1 de Rad53 : comparaison entre la structure complexée résolue par diffraction des rayons X et la structure non complexée observée par résonance magnétique nucléaire. **(A)** Poche de reconnaissance de l'aspartate dans le complexe FHA1 de Rad53 / peptide. Le domaine FHA est en vert et le fragment protéique en blanc (carbones), bleu (azotes), rouge (oxygènes) et orange (phosphate). L'arginine 83 du domaine FHA forme un pont salin avec l'aspartate du peptide en position pT+3, et trois autres liaisons hydrogènes avec des accepteurs du domaine FHA. Les liaisons hydrogènes sont représentées en pointillés bleus. **(B)** Poche de reconnaissance de l'aspartate lorsque le domaine n'est pas complexé. Le code des couleurs est le même que précédemment. On constate que l'arginine 83 est exposée au solvant. **(C)** Vue schématique du réseau de liaisons hydrogènes dans la poche de reconnaissance autour de la position pT+3 dans le complexe FHA1 de Rad53/peptide. Seules les liaisons mettant en jeu l'aspartate 8 et l'arginine 83 sont représentées. Les résidus du peptide sont représentés par des ovales gris, ceux du domaine FHA par des rectangles verts. Les liaisons hydrogènes sont indiquées par des pointillés bleus.

6.2 Mise en évidence du problème de l'exploration conformationnelle.

6.2.1 Introduction et description du système initial.

La structure cristallographique du complexe entre le domaine FHA1 de Rad53 et son peptide de plus haute affinité pTxxD est connue. Néanmoins, puisque nous souhaitons traiter le cas des structures pour lesquelles le criblage *in silico* doit tenir compte des mouvements du squelette, nous avons sélectionné comme structure de départ une structure du domaine FHA N-terminal de Rad53 non complexé obtenue par Résonance Magnétique Nucléaire sur laquelle le peptide natif a été amarré. Le *rmsd* séparant la structure du complexe natif et la structure du complexe ainsi recrée est de 1,42 Å au niveau de la zone de l'interface.

La séquence du peptide cristallographique, SLEV(pT)EAD, a été remplacée par une séquence poly-alanine AAAA(pT)AAA et la structure du complexe obtenu a été minimisée. Sur la base de cette structure de complexe, la position pT+3 est mutée en un des 20 acides aminés. Il ne s'agit pas de trouver l'orientation optimale du rotamère mais simplement de construire le peptide contenant la mutation : c'est la raison pour laquelle dans le cas de l'aspartate le rotamère initial pointe vers le solvant.

6.2.2 Contraintes relatives au positionnement conservé du peptide.

Dans les complexes FHA/peptides, le positionnement du peptide est bien conservé. D'un point de vue structural, ceci ce traduit notamment par la conservation de certaines liaisons hydrogènes entre d'un côté le squelette ou les chaînes latérales des résidus parfaitement conservés du domaine FHA et d'un autre côté le squelette ou le groupement phosphate de la thréonine du peptide phosphorylé. Ces différents contacts conservés sont représentés schématiquement au sein de la **figure 66** (traits gras). On note que les résidus en interaction avec le peptide phosphorylé font partie de trois boucles du domaine FHA. Quelques contacts intra-moléculaires conservés entre les résidus de ces boucles stabilisent la structure **figure 66** (traits fins).

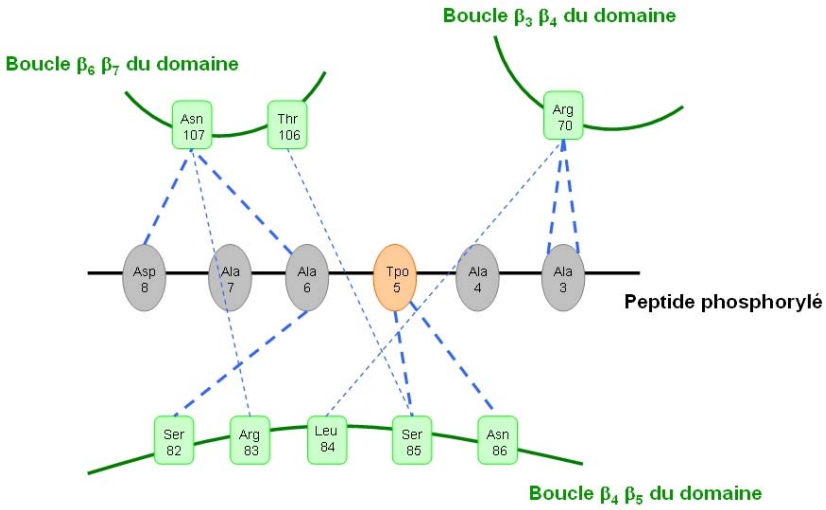


figure 66 : Représentation Schématique des contacts à l'interface du domaine FHA1 de Rad53 et du fragment phosphorylé conservés dans toutes les structures connues de complexe FHA/peptide. Les contacts intra-moléculaires sont en traits bleus fins et les contacts mettant en jeu le fragment phosphorylé sont en traits bleus gras.

Ainsi nous avons inclus des contraintes dans nos simulations de dynamique moléculaire pour que ces liaisons visant à positionner correctement le peptide soient formées. Dans la suite de ce manuscrit, ces contraintes seront appelées « contraintes d'homologie ».

6.2.3 Simulation longue de 5ns dans une boite d'eau.

La structure reconstruite à partir de la forme non liée du domaine FHA N-terminal de Rad53 complexée à un peptide pTxxD a constituée notre structure initiale. Au sein de ce complexe « hybride », l'aspartate en position pT+3 qui devrait être reconnue par l'arginine 83 est solvaté.

Afin de tester la capacité des approches de simulation de dynamique moléculaire à recréer le réseau de liaisons hydrogènes entre l'arginine 83 et le peptide caractérisé au sein de la structure native, chacune de ces liaisons a été analysée au cours de la dynamique. Les résultats sont présentés dans le tableau ci-dessous (**table 21**).

Nombre d'occurrences des liaisons hydrogènes impliquant l'arginine 83 et ses partenaires	Total	Pourcentage
Aucun contact natif	525	21,00
Un seul contact natif	519	20,76
Deux contacts natifs simultanés	895	35,80
Trois contacts natifs simultanés	543	21,72
Quatre contacts natifs simultanés	19	0,76
Cinq contacts natifs simultanés (topologie native)	0	0,00

table 21 : Table comptabilisant le nombre d'occurrences où les contacts natifs de l'arginine 83 et ses partenaires sont respectés. Le temps total de dynamique est de 5ns et une structure est analysée toutes les 2ps, ce qui ramène à 2500 le nombre de structures analysées.

La somme des différents pourcentages montre que les structures présentant deux contacts natifs ou moins représentent plus de 75% des structures produites par la simulation de dynamique moléculaire ! Les résultats montrent également que la topologie native avec les cinq liaisons hydrogènes formées n'est jamais explorée au cours des 5ns de dynamique.

Ces résultats suggèrent que la génération simultanée de tous les contacts natifs nécessaires à la reconnaissance de l'aspartate du peptide est un évènement extrêmement rare. Ce cas de figure n'est pas une exception. Le réseau d'interaction qui confère de la spécificité à une interaction suppose fréquemment qu'un couplage entre les différentes interactions favorables se produise. Par exemple, lorsqu'un résidu polaire se trouve enfoui dans une structure, la pénalité associée à sa désolvatation est tellement importante qu'elle impose que l'ensemble des groupements donneurs et accepteurs de liaisons hydrogènes soient satisfaits simultanément par des interactions dans cette structure. Dans la suite de ce travail exploratoire, nous introduisons une méthodologie basée sur l'utilisation de contraintes ambiguës dont l'objectif est de favoriser l'exploration d'états difficilement atteignables par l'approche de dynamique classique.

6.3 Utilisation de contraintes ambiguës.

6.3.1 Principe des contraintes ambiguës.

Pour diriger la dynamique, nous avons exploré la possibilité d'exploiter la complémentarité des interactions généralement observées dans les interfaces entre chaînes latérales des résidus (**figure 67**). Ainsi, les résidus hydrophobes vont tendre à s'amasser, et les atomes donneurs/accepteurs des résidus polaires et chargés vont chercher dans leur voisinage un atome complémentaire avec qui former une liaison hydrogène (**Annexe E**).

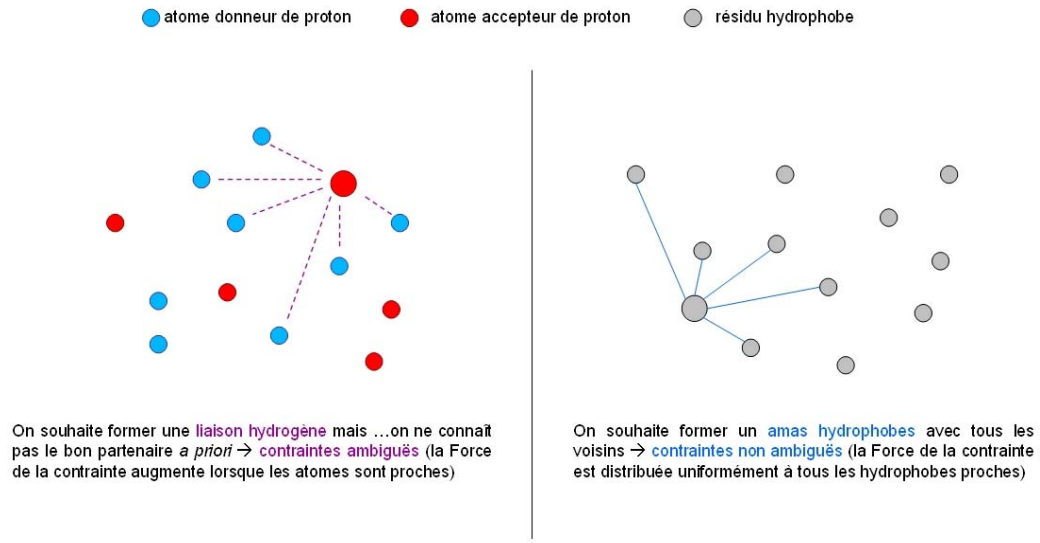


figure 67 : Principe de l'algorithme basé sur l'utilisation des contraintes en vue de diriger la dynamique. On distingue deux types de contraintes : celles visant à former des réseaux de liaisons hydrogènes et celles visant à former des amas hydrophobes.

L'insertion de ces contraintes donneurs/accepteurs se fait en deux étapes successives : dans une première phase, les contraintes sont appliquées aux atomes des chaînes latérales ; puis dans un second temps, les donneurs et les accepteurs du squelette peptidique sont à leur tour pris en compte. L'intérêt de cette dichotomie est d'utiliser la première étape pour induire des mouvements larges, notamment au sein des régions non structurées du domaine FHA qui sont au contact du peptide ; et d'utiliser la seconde étape pour améliorer la satisfaction des liaisons pour l'ensemble des atomes polaires enfouis par les contraintes.

Les contraintes imposées prennent la forme de contraintes ambiguës inspirées de celles utilisées pour résoudre les strutures par RMN. Ainsi, un ensemble de contraintes ambiguës

est imposé à chaque atome, et lorsque l'une d'entre elles est satisfaite, les autres sont relâchées. Insistons sur la notion qu'au sein de cet ensemble de contraintes, seule une fraction de l'ordre de 5% correspond à des contacts natifs.

6.3.2 Optimisation du problème de satisfaction de contraintes.

Avec cette méthode de génération de contraintes ambiguës/non-ambiguës se pose un réel problème d'optimisation. En effet, il faut que le système cherche et trouve la combinaison de contraintes la plus favorable, sachant que seule une faible proportion de contraintes imposées sont respectées au sein de la structure native. Sachant que la simulation de dynamique moléculaire n'est capable d'identifier que les *minima* locaux, il est préférable de relancer plusieurs fois le processus de simulation avec des conditions initiales variables.

Nous avons donc lancé 30 simulations de 150 ps chacune : 50 ps où ne sont inclues que les contraintes visant à positionner le peptide, 50 ps où sont inclues les contraintes ambiguës mettant en jeu les atomes des chaînes latérales, et 50 ps où sont inclues toutes les contraintes ambiguës.

6.3.3 Réduction du nombre d'atomes du système.

Deux stratégies ont été mises en place dans le but de réduire le temps de calcul nécessaire aux simulations de dynamique moléculaires. Tout d'abord, les mouvements des atomes ne sont calculés que pour la région au voisinage de l'interface FHA1 de Rad53/peptide : tous les résidus de la protéine dans une sphère de 15Å autour de la position pT+3 sont flexibles, les autres résidus du domaine étant bloqués au cours de la dynamique. Ensuite, pour économiser le temps de calcul tout en tirant les bénéfices d'une solvatation explicite, la solvatation est concentrée dans une bulle d'eau autour de l'interface FHA1 de Rad53/peptide. Cette bulle d'eau est maintenue stable grâce à une couche d'eau extérieure au seinde laquelle la position des molécules est restreinte par des contraintes harmoniques. La force des contraintes ainsi que le diamètre de la bulle ont été paramétrés de telle sorte que l'amplitude des mouvements des atomes de l'interface soit équivalente à celle observée dans la boite d'eau canonique utilisée pour la simulation de 5ns. La **figure 68** présente un aperçu du système final.

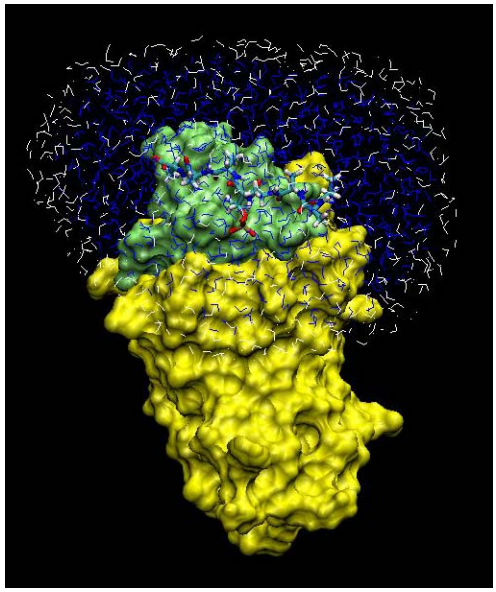


figure 68 : Le domaine FHA1 de Rad53 se décompose en une partie éloignée de la région de l'interface (en jaune) qui sera gelée, et une région proche de l'interface (en vert) flexible. Une bulle d'eau de 18 Å de rayon entoure la région flexible (en bleu, pour les molécules d'eau flexibles, en blanc pour les molécules d'eau qui constituent la couche d'étanchéité).

6.4 Application des contraintes ambiguës et solvatation dans une bulle d'eau.

6.4.1 Premiers résultats lors de la simulation du domaine FHA1 de Rad53 complexé à un fragment pTxxD.

Dans le cadre des 30 simulations lancées sur le domaine FHA1 de Rad53 complexé à un peptide contenant une aspartate en position pT+3. Une analyse des différents contacts du réseau de liaisons hydrogènes a été effectuée comme précédemment (**table 22**). Les résultats sont surprenants. D'un côté, on constate que le nombre de structures présentant un, deux ou trois contacts natifs baisse drastiquement. En revanche, alors que précédemment aucune structure échantillonnée ne présentait les cinq contacts natifs simultanément, l'application des contraintes ambiguës permet d'atteindre cette conformation (0,36% des structures).

Une inspection visuelle des structures échantillonnées au cours des 30 dynamiques permet de comprendre ce paradoxe. La **figure 69** présente certaines des structures les plus fréquemment générées. On constate que parmi ces structures, un certain nombre présentent des conformations inexactes dues à l'application des contraintes ambiguës (**figure 69-1,2,3**). Notamment, la présence d'un glutamate en position 137 piège fréquemment les dynamiques dans des conformations qui favorisent des liaisons hydrogènes entre les chaînes latérales de ce glutamate et celles de l'arginine 83.

Nombre d'occurrences des liaisons hydrogènes impliquant l'arginine 83 et ses partenaires (sur 4500)	Contraintes ambiguës dans l'eau (300K)	
	Total	Pourcentage
Aucun contact natif	4455	99,00
Un seul contact natif	0	0,00
Deux contacts natifs simultanés	1	0,02
Trois contacts natifs simultanés	37	0,82
Quatre contacts natifs simultanés	42	0,93
Cinq contacts natifs simultanés (topologie native)	16	0,36

table 22 : Table comptabilisant le nombre d'occurrences où les contacts natifs de l'arginine 83 et ses partenaires sont respectés. Le temps total de dynamique est de 5ns et une structure est analysée toutes les 2ps, ce qui ramène à 2500 le nombre de structures analysées.

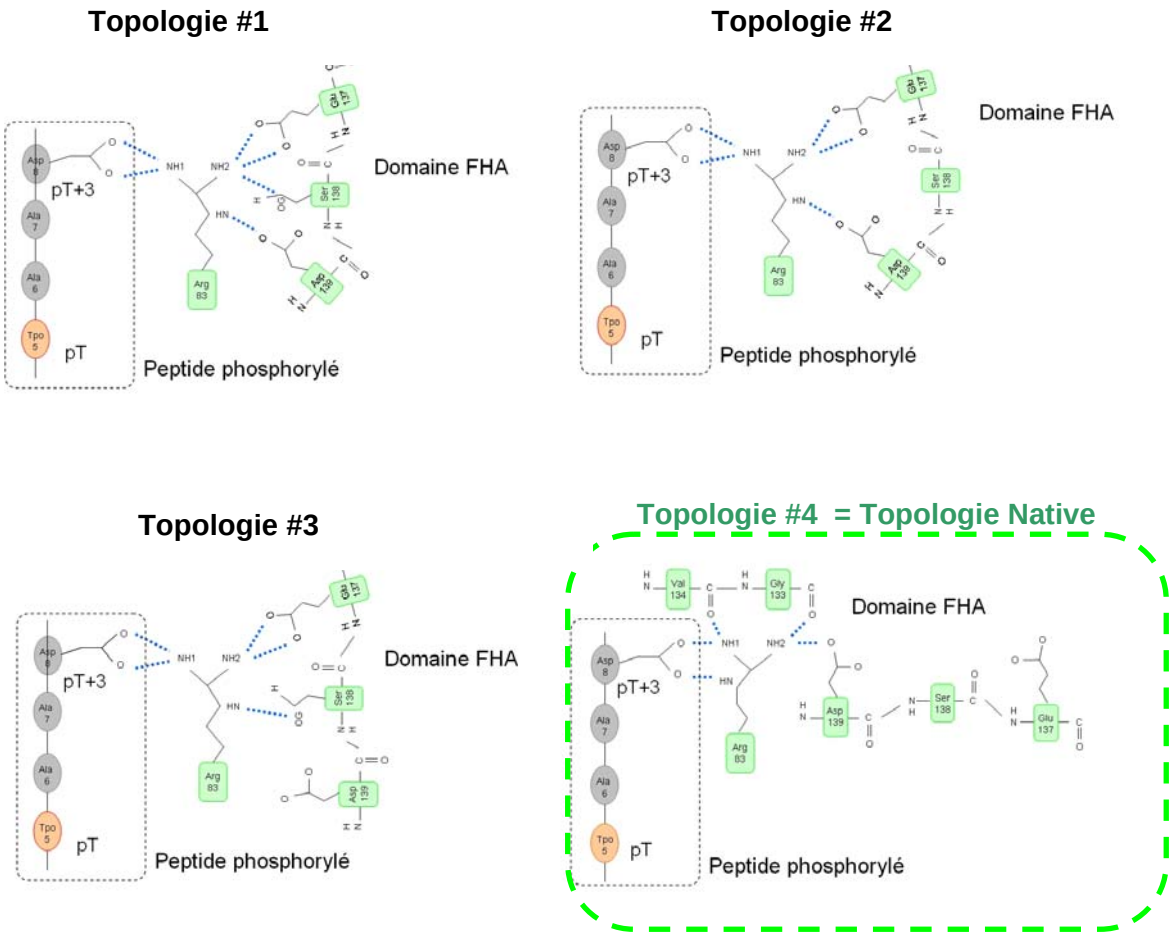


figure 69 : Réseaux de liaisons hydrogènes mis en place dans la poche de reconnaissance lorsque les contraintes ambiguës sont appliquées. Les simulations sont effectuées dans le champs de force OPLS de Gromacs, à une température de 300K. Les topologies exposées sont classées en fonction de leur fréquence d'apparition. Parmi les topologies explorées se trouve la topologie native (topologie #4).

Puisque la conformation native du réseau de liaison hydrogène est générée au cours de la dynamique sous contraintes ambiguës, deux questions nous ont intéressées et seront l'objet des paragraphes suivants.

(i) Une fonction d'énergie de type FOLDEF est-elle capable de discriminer et d'identifier ces structures aux contacts natifs parmi les différentes structures générées lors des 30 simulations du domaine FHA1 de Rad53 en contact avec le peptide pTxxD ?

(ii) Si tel est le cas, cette stratégie est-elle efficace dans le cadre d'un criblage pour identifier le motif pTxxD comme le plus affin ?

6.4.2 Première question : peut-on discriminer la conformation native du réseau de liaisons hydrogènes à l'aide d'une fonction d'évaluation ?

Au vu des résultats obtenus au chapitre précédent concernant le criblage virtuel de domaines FHA et BRCT sur squelette rigide, nous avons sélectionné la fonction d'énergie FOLDEF pour l'évaluation des structures échantillonnées par les simulations de dynamique moléculaires sous contraintes ambiguës.

Les résultats obtenus sont présentés sur la **figure 70**. Les structures respectant la topologie native du réseau de liaisons hydrogènes sont prédites comme ayant une énergie plus faible que les autres structures générées, ce qui traduit une stabilisation. Plus précisément, l'énergie libre du complexe est estimée à moins de 50 $kCal.mol^{-1}$ lorsque cette topologie est native, alors que ce seuil n'est pas franchi lorsque les topologies alternatives sont explorées.

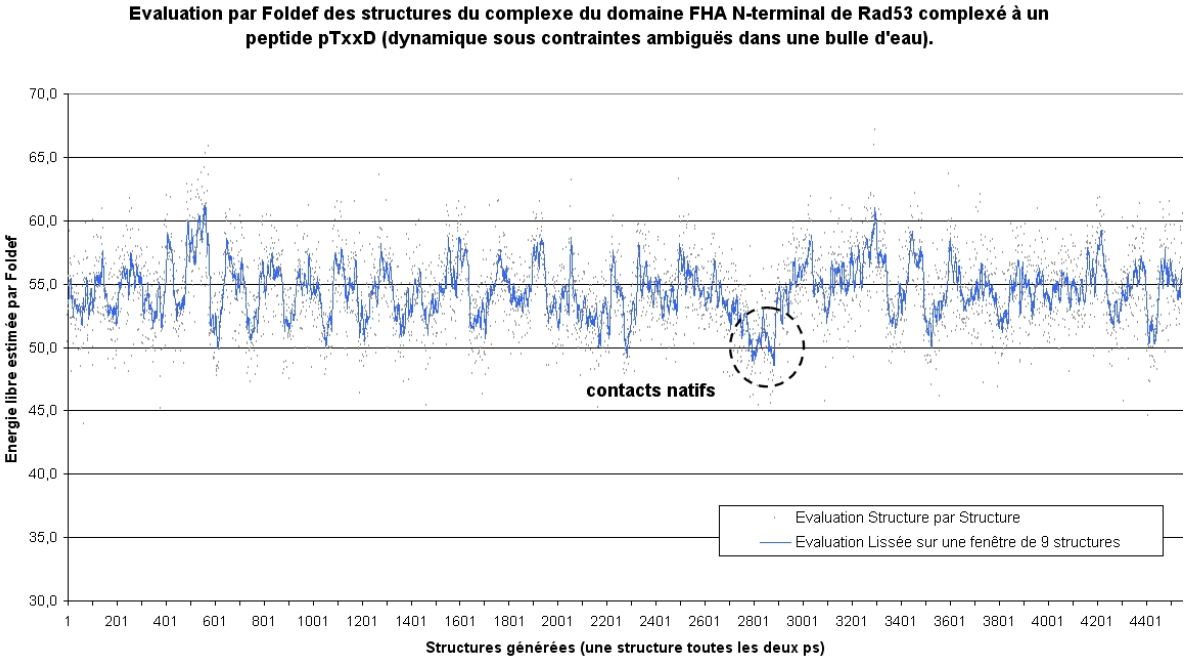


figure 70 : Evaluation des structures générées au cours de la dynamique par la fonction d'énergie Foldef. Pour plus de clarté, la courbe est lissée avec une fenêtre de 9 structures (4 structures en amont et 4 en aval).

On constate donc que la topologie native est correctement discriminée : les différences d'énergie libre entre ces différentes structures sont faibles mais significatives au regard de la marge d'erreur de la fonction d'énergie FOLDEF (0,80 kCal.mol^{-1}).

6.4.3 Deuxième question : peut-on appliquer cette stratégie dans le cadre d'un criblage *in silico* ?

La stratégie utilisée pour générer les 30 dynamiques du domaine FHA N-terminal de Rad53 complexée à un peptide pTxxD a été repétée pour chacun des 19 acides aminés substitués en position pT+3, en appliquant le même algorithme de génération de contraintes ambiguës. Cette opération coûteuse en temps n'aurait pas été réaliste sans que la protéine ne soit rigidifiée sur les régions éloignées de l'interface et sans que la solvatation explicite ne soit restreinte à cette même région. Les dynamiques obtenues ont été évaluées avec la fonction d'énergie FOLDEF. La **table 23** présente les résultats complets de cette évaluation.

motif criblé in silico	score Foldef de la meilleure des 30 simulations	ΔΔG avec le motif pTxxA
pTxxA	47,90	0,00
pTxxC	49,10	1,20
pTxxD	49,96	2,06
pTxxE	47,55	-0,35
pTxxF	49,02	1,12
pTxxG	49,28	1,38
pTxxH	47,51	-0,39
pTxxI	47,53	-0,37
pTxxK	49,61	1,71
pTxxL	47,73	-0,17
pTxxM	49,84	1,94
pTxxN	49,02	1,12
pTxxP	50,31	2,41
pTxxQ	47,59	-0,31
pTxxR	49,77	1,87
pTxxS	48,14	0,24
pTxxT	49,47	1,57
pTxxV	50,70	2,80
pTxxW	48,84	0,94
pTxxY	47,57	-0,33

table 23 : Résultat du criblage virtuel des différents phospho-peptides. L'exploration conformationnelle a été effectuée par simulation de dynamique moléculaire, sous contraintes ambiguës et dans une bulle d'eau. Les structures générées au cours de ces différentes dynamiques ont été évaluées par la fonction d'évaluation Foldef. Pour chaque phospho-peptide criblé, 30 simulations ont été effectuées. Les résultats présentés ici sont ceux de la meilleure dynamique.

Les résultats obtenus sont décevants puisqu'on constate que la quasi-totalité des autres peptides sont prédits comme plus affins que le peptide pTxxD. Seuls deux peptides sont estimés par Foldef comme moins favorables : il s'agit des peptides pTxxP (50,31 $kCal.mol^{-1}$) et pTxxV (50,70 $kCal.mol^{-1}$).

Ce résultat peut s'expliquer par le fait que de multiples interactions à l'interface entre le domaine FHA et son peptide contribuent à la valeur de l'énergie estimée. Dans le cas du peptide TxxD, la structure de plus basse énergie correspond bien à l'interaction native entre l'aspartate en position pT+3 et l'arginine 83. Néanmoins, cette conformation a été rarement explorée et il est possible que dans les autres zones de cette interface, des interactions défavorables aient contribué à réduire la stabilité. Pour pallier à ce problème, il faudrait explorer de façon approfondie les interfaces pré-sélectionnées lors d'un premier crible comme celui présenté **figure 70**. En termes de temps de calcul, cette solution risque de devenir rapidement prohibitive car elle suppose un processus itératif dans lequel le poids de certaines contraintes ambiguës évoluerait pour favoriser/défavoriser certaines contraintes.

Cette stratégie utilisant des contraintes ambiguës dont la force est modulée itérativement en fonction de l'énergie des modèles générés à l'itération précédente est désormais couramment utilisée dans les programmes de résolution des structures par RMN (Linge et al., 2003). Néanmoins, pour que cette approche soit envisageable, il sera nécessaire d'échantillonner plus efficacement les solutions proches de la conformation native.

6.5 Conclusions et perspectives

La prise en compte de la flexibilité du squelette peptidique dans la modélisation des interfaces protéine-protéine ou protéine-ligand constitue un enjeu majeur pour les méthodes de criblage *in silico*. Dans ce chapitre, nous avons focalisé la présentation des résultats sur une perspective qui nous apparaît très prometteuse.

Avant d'utiliser la simulation de dynamique moléculaire pour rendre compte de la flexibilité, nous avons testé de nombreuses alternatives telles que la modélisation des boucles par des programmes tels que MODELLER (Sali and Blundell, 1993) ou RAPPER (DePristo et al., 2003). Les résultats obtenus n'ont jamais permis d'atteindre une précision suffisante (données non exposées). Les principaux défauts de ces approches étaient : (i) un champ de force inadapté pour générer des conformations vraisemblables dans MODELLER ; (ii) un échantillonnage trop réduit des conformations explorées par l'algorithme Monte-Carlo de RAPPER. De ce point de vue, l'utilisation du champ de force OPLS, de la solvatation explicite et du programme de simulation Gromacs a permis d'améliorer sensiblement la qualité des structures générées, au détriment du temps de calcul.

Un des problèmes fondamentaux de ces outils de simulation est que le système se trouve facilement piégé dans des états d'énergie localement favorables qui limitent l'exploration exhaustive du paysage énergétique. Pour dépasser cette limitation, nous avons étudié la pertinence de l'introduction de contraintes ambiguës pour générer et identifier la structure présentant un ensemble d'interactions optimales entre atomes d'une interface. L'intérêt et l'originalité de cette stratégie est que ce sont les propriétés des chaînes latérales qui guident les changements de conformations du squelette et non, comme dans les méthodes de modélisation des boucles, l'échantillonnage du squelette peptidique qui guide le

positionnement des chaînes latérales. Sur l'exemple étudié, l'introduction de contraintes ambiguës basées sur la complémentarité physico-chimique des chaînes latérales permet de générer un état rare et difficilement atteignable sans l'ajout de contraintes. Il est de plus intéressant de noter que lorsque cette structure est générée, une fonction d'évaluation du type de FOLDEF permet de la sélectionner parmi les différentes structures échantillonnées.

Afin de poursuivre ce travail, il sera important d'appliquer cette méthodologie couplant flexibilité et contraintes ambiguës à d'autres interfaces. A ce titre, les domaines FHA constituent des exemples de choix en raison de leur grande versatilité de reconnaissance. On pourra ainsi tester si les contraintes ambiguës permettent de prédire avec la même efficacité des interfaces basées sur des contacts hydrophobes.

Conclusions et Perspectives

Les réseaux d'interactions protéine-protéine mis en place au sein de la cellule dans le but de transmettre un signal ou de répondre à celui-ci sont complexes et l'identification du rôle précis de chaque interaction au sein du réseau est un enjeu actuel majeur. La régulation de ces interactions au niveau intra- et inter-moléculaire par des mécanismes d'allostérie constitue un second thème de recherche porteur. Les domaines chargés d'opérer cette régulation fine ont des propriétés surprenantes et leur étude est délicate. Un des objectifs majeurs de ce travail de thèse est de mieux caractériser les propriétés des modules de reconnaissance peptidique pour prédire efficacement leurs sites de liaisons.

Le premier point abordé pendant cette thèse a été la modélisation structurale des PRMs. Plus particulièrement, la protéine humaine Nbs1 et son orthologue Xrs2 chez *Saccharomyces cerevisiae* ont été étudiées. Suite à la découverte d'un tandem de domaines BRCT au sein de ces deux protéines dans une région extrêmement divergente, nous avons réalisé un modèle structural validé par des fonctions d'évaluation classiques et par simulation de dynamique moléculaire. La mise en évidence de ce tandem domaines BRCT ouvre la voie à de nouvelles hypothèse concernant les interactions induites par une cassure double brin de l'ADN et leur lien avec le syndrome de Nijmengen.

La modélisation structurale des tandems de domaines BRCT des protéines Nbs1 et Xrs2 nous a confronté à un problème de modélisation par bioinformatique de ces courts domaines hautement divergents. En effet, les méthodes de modélisation actuelles reposent toutes sur la production d'un alignement de haute qualité qui sert de base à une reconstruction par homologie. Malheureusement, cet alignement est difficile à mettre au point lorsque les séquences partageant un même repliement sont très divergentes. Nous avons développé une approche permettant d'explorer de façon ciblée l'ensemble des meilleurs alignements. Cette méthode s'avère utile lorsque l'alignement optimal du point de vue des séquences ne permet pas de générer une structure de qualité satisfaisante. En nous plaçant au sein du formalisme très général et rigoureux des modèles de Markov cachés, nous avons pu profiter d'un algorithme existant et l'implémenter comme une nouvelle fonctionnalité au sein du programme HMMer. L'efficacité de cette méthode pour générer des alignements plus fiables a été mise en évidence par une étude comparative à grande échelle. Une extension naturelle de ce travail consistera à implémenter ce même algorithme dans le cadre des méthodes

d'alignements profil-profil utilisant elles aussi des modèles de Markov cachés (PRC ou HHPred).

L'obtention d'un modèle n'est pas une fin en soi. Idéalement, on souhaiterait sur la base de ces modèles prédire les propriétés fonctionnelles de ces domaines. Dans le cas des PRMs, l'une des interrogations fondamentales est liée au motif linéaire qu'ils sont capables de reconnaître de façon sélective. Le chapitre 4 illustre l'intérêt de connaître la spécificité de liaison pour identifier efficacement les sites d'interactions entre partenaires. La stratégie employée, élaborée dans le cadre du projet SpIDER, a déjà permis d'identifier six sites d'interactions entre la kinase Rad53 et ses partenaires. Sans l'apport de la prédiction bioinformatique, cette démarche se serait révélée extrêmement aléatoire et fastidieuse. Le cas de l'interaction entre Rad53 et Cdc45 est à ce titre exemplaire. Cdc45 étant un gène essentiel, il aurait été difficile par des cribles de génétique fonctionnelle (recherche d'épistasie par double délétion) d'identifier le rôle de cette interaction. Dans notre cas, il s'agit d'un candidat d'intérêt puisque le mutant ponctuel de Cdc45 qui perd l'interaction avec Rad53 possède des phénotypes remarquables suite à un stress génotoxique. L'ensemble de ce travail souligne l'importance des développements méthodologiques en bioinformatique visant à prédire les spécificités de liaison des PRMs.

Afin d'étendre ce type d'analyse à d'autres domaines PRMs, nous avons recherché comment prédire par criblage *in silico* les motifs spécifiquement reconnus, en s'appuyant sur les exemples des domaines FHA et BRCT. Deux cas de figure ont été distingués : (i) la structure criblée peut être maintenue rigide au niveau du squelette peptidique ; (ii) la structure criblée nécessite que le squelette peptidique soit rendu flexible au niveau de l'interface. Le premier cas peut s'appliquer si on possède une structure complexée PRM-ligand ou si la structure du PRM libre est peu alterée par l'interaction avec le ligand. Le second constitue une généralisation de l'approche de criblage qui pourra s'appliquer aussi bien aux structures expérimentales qu'aux structures modélisées.

Dans le premier cas de figure, nous avons montré comment les méthodes de *design* pouvaient être détournées de leur objectif initial pour prédire des motifs consensus de liaison. Les méthodes ROSETTA et FOLDEF, en plus d'être rapide, donnent dans la majorité des exemples traités des prédictions en adéquation avec les observations expérimentales.

Néanmoins, nous avons mis en évidence des problèmes de sélectivité et proposé des pistes pour améliorer la fiabilité des prédictions. En particulier, nous pensons qu'il est important d'intégrer la notion de « *design* négatif » dans la phase d'apprentissage de ces fonctions d'énergie afin de réduire la proportion de faux positifs détectés.

La version du programme FOLD-X (calculant la fonction d'énergie FOLDEF) que nous avons utilisée ne possédait pas de moteur d'exploration conformationnelle permettant d'optimiser le placement des chaînes latérales. La qualité des modèles obtenus avec ROSETTA confirme que le couplage entre optimisation et évaluation est plus efficace que les stratégies où les deux problèmes sont traités indépendamment. Dans la prochaine version de FOLD-X, cette fonctionnalité sera disponible (communication personnelle de François Stricher et Luis Serrano).

Dans la dernière partie de cette thèse, nous avons abordé le problème de la prédiction des spécificités de liaison des domaines de reconnaissance peptidique lorsque le squelette de l'interface nécessite d'être rendu flexible. Ce cas de figure est important à considérer. En effet, de nombreuses structures de PRMs ont été résolues sous leur forme non-complexée dans le cadre de projets de génomique structurale. De plus, au cours des années à venir, il est certain que la majorité des structures des PRMs pourra être modélisée par des approches bioinformatiques. Cela signifie qu'il est capital de mettre au point des approches de criblage *in silico* capables de se satisfaire de structures initiales dont le squelette est susceptible de bouger. Au cours de cette thèse, nous avons testée des approches à base de simulation de dynamique moléculaire pour prendre en compte ces mouvements. En guidant la dynamique par un ensemble de contraintes ambiguës inspirées de celles utilisées pour résoudre les structures par RMN, nous avons obtenu des résultats tout à fait prometteurs. Néanmoins, cette stratégie est coûteuse en temps de calcul ce qui limite son applicabilité. Les efforts doivent donc selon nous se concentrer sur la prise en compte des mouvements du squelette par des méthodes moins gourmandes en temps de calcul.

Parmi les pistes à explorer pour résoudre ce problème, nous pensons que l'introduction de la flexibilité au sein des approches discrètes FOLDEF et ROSETTA est la plus prometteuse. La suite ROSETTA intègre depuis peu un moteur autorisant de nombreux types de mouvements au niveau du squelette : (i) des mouvements de faible amplitude couplés à une description

atomique du système ; (ii) des mouvements de plus grande amplitude associés à une représentation simplifiée en particulier au niveau des chaînes latérales. L'introduction de ces degrés de liberté supplémentaires rend le problème de l'exploration de l'espace conformationnel central. L'algorithme utilisé dans ROSETTA pour l'exploration de cet espace est basé sur une optimisation de type Monte-Carlo coûteuse en temps. A titre d'exemple, les calculs effectués par l'équipe de D. Baker reposent sur l'utilisation de calculs distribués sur une vaste grille. D'autres algorithmes prometteurs pour gérer la flexibilité du squelette peptidique inspirés des approches de robotique ont été proposés (Cortes et al., 2005). Dans l'avenir, il sera important d'évaluer si ces stratégies alternatives, couplées à des approches du type FOLDEF ou ROSETTA, améliorent le pouvoir prédictif des méthodes bioinformatiques. Ces améliorations sont en effet cruciales pour rendre compte de la complexité des interactions atomiques qui assurent la spécificité de la reconnaissance entre PRM et ligand et dont nous avons pu apprécier le raffinement au cours de cette thèse.

Annexes : Matériel et Méthodes

A. Mise en place de la fonction HmmKalign au sein de HMMer.

- ✓ Utilisation de la commande HmmKalign.

Nous avons introduit une nouvelle fonction, HmmKalign, au sein de HMMer (**Article 2**). Pour utiliser cette commande, il faut avoir préalablement construit un modèle de Markov caché, ce qui ce fait classiquement grâce à la commande :

```
$ hmmbuild [-options] <hmm file> <multiple alignment file>
```

Ensuite, pour aligner une séquence s sur le modèle de Markov caché construit, il faut créer un fichier `<two sequences file>` contenant deux séquences au format fasta dans l'ordre suivant :

1. la séquence s ;
2. l'une des séquences présentes dans `<multiple alignment file>` ; typiquement, lorsque l'on souhaite faire un modèle par homologie de la séquence s, cette seconde séquence est celle correspondant à la structure modèle.

Pour obtenir les k meilleurs alignements de s sur le modèle de Markov caché construit, on utilise la commande :

```
$ hmmkalign k <hmm file> <two sequences file>
```

- ✓ Composition de la base de test.

L'ensemble des 23 alignements structuraux utilisé pour l'étude est dérivé de la base de données HOMSTRAD (Mizuguchi et al., 1998; Stebbings and Mizuguchi, 2004). Les familles ont été sélectionnées en fonction des critères suivants :

(i) la famille regroupe au moins cinq structures expérimentales ;

(ii) l'identité de séquence moyenne au sein de la famille est inférieure ou égale à 25%.

Les noms HOMSTRAD des 23 familles retenues au sein de la base d'étude sont : ABC transporter (identité de séquence moyenne 25%), acetyltransferase family (24%), alpha

amylase, C-terminal domain (21%), alpha amylase, catalytic domain (23%), anticodon binding domain (22%), cytochrome p450 (21%), DEATH domain (20%), fibronectin type III domain (16%), glycosyl hy-drolase family 5 (18%), haloperoxidase (25%), histidine kinase, DNA gyrase B and HSP90-like ATPase (25%), integrin I-domain (22%), kinase (25%), lipo-calin family (20%), metallo-beta-lactamase superfamily (21%), PH domain (17%), proteasome A-type and B-type (22%), reductases (22%), rhodanese-like domain (23%), RNA recognition motif (23%), short-chain dehydro-genases/reductases (23%), thioredoxin (20%), TPR domain (17%).

Au sein de chaque famille, 5 séquences s_{obs} ont été sélectionnées aléatoirement. Pour chaque séquence, le HMM décrivant la famille a été généré à partir de l'alignement structural de tous les membres de la famille à l'exception de s_{obs} et des séquences partageant plus de 40% d'identité avec s_{obs} (afin d'éviter des biais trop favorables). Le fait de sélectionner exactement le même nombre de séquences par famille permet d'éviter la sur-représentation des familles les plus peuplées.

- ✓ Construction des HMM : méthode utilisant la conservation des structures secondaires.

Le fait de considérer une colonne de l'alignement multiple comme un état d'appariement ou comme une insertions entre deux états d'appariement consécutifs est un choix crucial dans le cadre de la génération des κ meilleurs alignements.

Construire l'architecture du modèle de Markov caché en se basant sur la conservation des structures secondaires au sein de la famille consiste à restreindre les états d'appariements aux états qui respectent l'une des deux conditions suivantes (voir détails ci-dessous):

(i) la position fait partie d'une structure secondaire dans la majorité des séquences ;

(ii) la position est très conservée en terme de séquence.

Plus précisément, une position p est considérée comme un état d'appariement en raison de la conservation des structures secondaires lorsque :

$$\sum_{i\in\{1..k\}} ss(p,i)\times w(i) \geq 1/2$$

avec $s_1,s_2....s_k$ représentant l'ensemble des κ séquences, $w(i)$ représentant le poids GSC de la séquence s_i et $ss(p,i)$ étant un booléen égal à 1 si la position est une hélice ou un feuillet, et 0 sinon. Les annotations de structures secondaires sont directement extraites de la base de données HOMSTRAD, où elles ont été pré-calculées à l'aide de la méthode JOY (Mizuguchi et al., 1998).

Le programme RATE4SITE (Pupko et al., 2002) a été utilisé pour déterminer les positions très conservées en terme de séquence (seuil 0.6).

- ✓ Test de Student sur la moyenne des entropies, des Qmod Qdev et Qloc.

Afin de vérifier que la différence des moyennes est significative, on réalise un test de Student. Dans le cadre des grands nombres (n ≥ 30), ce test est robuste et applicable aux échantillons non-gaussiens. L'hypothèse $\mathcal{H}$: $m_1=m_2$ est testée, où m_1 et m_2 sont les moyennes des deux distributions.

On considère la variable :

$$T = \frac{M_1 - M_2 - (m_1 - m_2)}{\sqrt{\left(n_1 S_1^{\,2} + n_2 S_2^{\,2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sqrt{n_1 + n_2 - 2}$$

où n_1 et n_2 sont les tailles des deux échantillons (dans notre cas $n_1=n_2=n$ =115) ; m_1 et m_2 sont les espérances ; M_1 et M_2 sont les moyennes empiriques et S_1 et S_2 les écarts-types empiriques. Cette variable suit une loi de Student de paramètre $2n-2$. On a notamment :

P(|T|>1.96) = 0.05, ce qui correspond à un risque d'erreur de 5% ;

P(|T|>2.58) = 0.01, ce qui correspond à un risque d'erreur de 1% ;

P(|T|>3.29) = 0.001, ce qui correspond à un risque d'erreur de 0.1%.

Sous l'hypothèse $\mathcal{H}$, $T = \frac{M_1 - M_2}{\sqrt{\frac{S_1^{\,2} + S_2^{\,2}}{n-1}}}$.

Sur la totalité de la longueur des séquences, l'entropie moyenne des 20 alignements générés par hmmkalign est de :

- 0.1698 (écart type 0.0166) lorsque le HMM est construit traditionnellement ;
- 0.2539 (écart type 0.0205) lorsque le HMM est construit en restreignant les états d'appariements aux positions de structures secondaires conservées.

D'où :

T = (0.1698 -0.2539) / racine ((0.0166+0.0205)/114)

T = 0.0842 / racine (0.0371/114)

T = 0.0842 / 0.018039917

T = 4.6674273

L'hypothèse $\mathcal{H}$ est donc rejetée avec un risque de 0.1% d'erreur.

En restreignant les calculs aux régions de structures secondaires, l'entropie moyenne des 20 alignements générés par hmmkalign est de :

- 0.1239 (écart type 0.0179) lorsque le HMM est construit traditionnellement ;
- 0.1617 (écart type 0.0179) lorsque le HMM est construit en restreignant les états d'appariements aux positions de structures secondaires conservées.

D'où :

T = (0.1239 -0.1617) / racine ((0.0179+0.0179)/114)

T = 0.0378 / racine (0.0358/114)

T = 0.0378 / 0.01772

T = 2.13305

L'hypothèse $\mathcal{H}$ est donc rejetée avec un risque de 5% d'erreur.

Les tests de Student concernant les différences des moyennes sur le Q_{mod}, le Q_{dev} et le Q_{local} ont été effectués de façon similaire.

- ✓ Utilisation de HHPred pour générer des alignements profil-profil.

Nous avons utilisé HHPred (Soding, 2005) avec utilisation des prédictions de structures secondaires, car il a été montré que la méthode est plus fiable ainsi. Pour que la comparaison

avec les résultats de HmmKalign soit plus juste, nous avons utilisé les structures secondaires exactes (extraites de HOMSTRAD) et non pas prédites par un programme spécifique. En effet, dans notre procédure de test de HmmKalign, se sont les structures secondaires exactes qui ont servi à l'élaboration des HMMs basés sur la conservation des structures secondaires.

Construction du HMM de la famille. Les alignements structuraux ont été enrichis avec leurs homologues de plus de 40% obtenus par une recherche PSI-BLAST (sur la banque nr70). Les redondances ont été éliminées au sein du pool de séquences. L'alignement enrichi de ces homologues s'est ensuite effectué sans perturbation de l'alignement structural initial avec Clustal. Les structures secondaires exactes ont été utilisées pour annoter l'alignement obtenu : les séquences de l'alignement structural initial ont été pondérées par la méthode GSC et les annotations de la banque de données HOMSTRAD ont été extraites.

Construction du HMM de la séquence test. Pour la séquence test, la recherche des homologues de plus de 40% s'est effectuée par un Blast sur la banque nr70. L'alignement a ensuite été généré par Clustal et les structures secondaires ont été prédites par le programme Psi-Pred.

Parmi les 115 cas testés, 4 cas n'ont pas été traités par HHPred : il s'agit des séquences dont les codes PDB sont 3hhrb1, 1avaa, 1gjwak et 1esl. Lors du traitement de ces quatre séquences, les HMMs produits par HHPred étaient vides. Nous avons constaté que ces quatre cas correspondaient à des problèmes de redondance partielle de nom de fichiers.

B. Détection des sites reconnus par les domaines FHA de Rad53.

✓ Résultats obtenus pour l'interaction entre le domaine FHA1 de Rad53 et Ptc2.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
22	ADSLTAFGL	0.024	pTxxG	(7/7)	(7/7)
48	PNVLTKSDK	0.543	pTxxD	(7/7)	(5/7)
99	ALIDTFINT	0.059	pTxxN	(7/7)	(3/7)
103	TFINTDVKL	0.200	pTxxK	(7/7)	(6/7)
121	HSGCTATSI	0.015	pTxxS	(7/7)	(7/7)
123	GCTATSILV	0.067	pTxxL	(7/7)	(7/7)
144	GDSRTVLAT	0.019	pTxxA	(7/7)	(7/7)
148	TVLATDGNA	0.011	pTxxN	(4/7)	(7/7)
162	DHKPTLASE	0.101	pTxxS	(7/7)	(7/7)
211	EQIVTCVPD	0.196	pTxxP	(7/7)	(7/7)
241	WDCLTSQDC	0.095	pTxxD	(4/7)	(7/7)
258	REGKTLNEI	0.252	pTxxE	(7/7)	(7/7)
274	CCAPTTEGT	0.229	pTxxG	(7/7)	(7/7)
275	CAPTTEGTG	0.107	pTxxT	(7/7)	(7/7)
278	TTEGTGIGC	0.085	pTxxG	(7/7)	(7/7)
313	KAHRTSVRS	0.646	pTxxR	(5/7)	(7/7)
343	VFAITTKKP	0.855	pTxxK	(7/7)	(5/7)
344	FAITTKKPQ	0.938	pTxxP	(7/7)	(6/7)
352	QDKFTRDHE	0.032	pTxxH	(5/7)	(7/7)
363	VASVTAADN	0.038	pTxxD	(7/7)	(6/7)
376	**DIDDTDADT**	**0.642**	**pTxxD**	**(7/7)**	**(7/7)**
380	TDADTDAEN	0.136	pTxxE	(7/7)	(7/7)
395	SKSKTSGPI	0.059	pTxxP	(5/7)	(5/7)
411	LLGATGGVK	0.025	pTxxV	(7/7)	(6/7)
416	GGVKTDSNG	0.063	pTxxN	(6/7)	(5/7)
424	GNKVTYTLP	0.012	pTxxL	(7/7)	(7/7)
426	KVTYTLPQS	0.032	pTxxQ	(7/7)	(5/7)
438	QLLQTMGHD	0.026	pTxxH	(5/7)	(5/7)
454	NDSNTDHKA	0.068	pTxxK	(1/7)	(4/7)

✓ Résultats obtenus pour l'interaction entre le domaine FHA1 de Rad53 et Asf1.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
19	PAKFTDPYE	0.019	pTxxY	(9/9)	(9/9)
27	EFEITFECL	0.068	pTxxC	(9/9)	(9/9)
43	EWKLTYVGS	0.338	pTxxG	(9/9)	(9/9)
93	LVSVTVILL	0.007	pTxxL	(9/9)	(9/9)
147	KPRVTRFNI	0.224	pTxxN	(9/9)	(9/9)
215	EEEKTEDNE	0.818	pTxxN	(7/9)	(7/9)
220	EDNETNLEE	0.041	pTxxE	(7/9)	(8/9)
265	EGGSTDIES	0.262	pTxxE	(9/9)	(4/9)
270	**DIESTPKDA**	**0.994**	**pTxxD**	**(8/9)**	**(7/9)**

✓ Résultats obtenus pour l'interaction entre le domaine FHA1 de Rad53 et Cdc45.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
42	ALCATKMLS	0.021	pTxxL	(6/6)	(6/6)
107	YVIDTDEKS	0.478	pTxxK	(4/6)	(2/6)
147	FDDGTVDDT	0.681	pTxxD	(6/6)	(5/6)
151	TVDDTLGEQ	0.305	pTxxE	(4/6)	(6/6)
189	**DDEATDADE**	**0.339**	**pTxxD**	**(4/6)**	**(5/6)**
195	ADEVTDEDE	0.124	pTxxD	(3/6)	(6/6)
205	DEDETISNK	0.358	pTxxN	(1/6)	(2/6)
245	YSQGTTVVN	0.682	pTxxV	(5/6)	(5/6)
246	SQGTTVVNS	0.235	pTxxN	(6/6)	(6/6)
265	AIGETNLSN	0.007	pTxxS	(6/6)	(4/6)
277	NILGTTSLD	0.058	pTxxL	(2/6)	(6/6)
278	ILGTTSLDI	0.085	pTxxD	(5/6)	(6/6)
303	VKRLTPSSR	0.934	pTxxS	(4/6)	(2/6)
312	NSVKTPDTL	0.804	pTxxT	(6/6)	(6/6)
315	KTPDTLTLN	0.174	pTxxL	(6/6)	(6/6)
317	PDTLTLNIQ	0.019	pTxxI	(2/6)	(6/6)
372	IPLSTAQET	0.528	pTxxE	(5/6)	(6/6)
376	TAQETWLYM	0.008	pTxxY	(4/6)	(6/6)
412	GFVRTLGYR	0.030	pTxxY	(6/6)	(6/6)
429	VEALTALLE	0.071	pTxxL	(6/6)	(6/6)
438	VGNSTDKDS	0.631	pTxxD	(2/6)	(5/6)
453	NNDDTDGEE	0.573	pTxxE	(4/6)	(5/6)
467	AQKLTNLRK	0.009	pTxxR	(3/6)	(5/6)
509	AIFNTGVAI	0.022	pTxxA	(6/6)	(6/6)
544	RNPLTLLRL	0.071	pTxxR	(5/6)	(6/6)
576	IDENTDTYL	0.027	pTxxY	(5/6)	(6/6)
578	ENTDTYLVA	0.219	pTxxV	(6/6)	(6/6)
585	VAGLTPRYP	0.920	pTxxY	(4/6)	(6/6)
594	RGLDTIHTK	0.345	pTxxT	(6/6)	(5/6)
597	DTIHTKKPI	0.727	pTxxP	(5/6)	(6/6)
613	FQQITAETD	0.015	pTxxT	(5/6)	(6/6)
616	ITAETDAKV	0.246	pTxxK	(6/6)	(6/6)
645	LEKLTLSGL	0.084	pTxxG	(6/6)	(6/6)

✓ Résultats obtenus pour l'interaction entre le domaine FHA2 de Rad53 et Nse5.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
25	FVELTEKHL	0.774	pTxxH	(8/8)	(8/8)
74	LVLFTLSTL	0.008	pTxxT	(8/8)	(8/8)
77	FTLSTLSEY	0.272	pTxxE	(8/8)	(7/8)
95	DPYNTSRET	0.996	pTxxE	(1/8)	(5/8)
99	TSRETLSRR	0.570	pTxxR	(7/8)	(5/8)
141	LAIDTLTPK	0.527	pTxxP	(6/8)	(8/8)
143	IDTLTPKKQ	0.920	pTxxK	(6/8)	(6/8)
157	RFRRTKSES	0.965	pTxxE	(3/8)	(6/8)

164	ESGVTYRQN	0.430	pTxxQ	(3/8)	(5/8)
180	DQAKTFKNP	0.619	pTxxN	(7/8)	(8/8)
197	EQRNTILGN	0.422	pTxxG	(6/8)	(7/8)
221	MILWTLSNS	0.014	pTxxN	(7/8)	(7/8)
230	LQESTPLFL	0.067	pTxxF	(6/8)	(1/8)
290	ESLNTRNFA	0.025	pTxxF	(7/8)	(7/8)
316	DNYATPVHP	0.605	pTxxH	(7/8)	(8/8)
327	NGENTIVDT	0.984	pTxxD	(6/8)	(1/8)
331	TIVDTYIPT	0.266	pTxxP	(6/8)	(6/8)
335	TYIPTIKCS	0.367	pTxxC	(1/8)	(2/8)
373	**HRLITPRIV**	**0.984**	**pTxxI**	**(6/8)**	**(7/8)**
388	GISRTLASF	0.021	pTxxS	(7/8)	(7/8)
404	KFFMTENLS	0.092	pTxxL	(7/8)	(7/8)
421	LAEGTLSEI	0.543	pTxxE	(7/8)	(8/8)
429	ILKDTQECV	0.197	pTxxC	(6/8)	(6/8)
437	VVILTLVEN	0.647	pTxxE	(6/8)	(7/8)
464	CFAFTEQCS	0.025	pTxxC	(6/8)	(3/8)
499	HLIGTEAIL	0.052	pTxxI	(1/8)	(7/8)

✓ Résultats obtenus pour l'interaction entre le domaine FHA2 de Rad53 et Ste5.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
4	-MMETPTDN	0.795	pTxxD	(6/7)	(4/7)
6	METPTDNIV	0.020	pTxxI	(6/7)	(6/7)
20	FGSSTQYSG	0.157	pTxxS	(4/7)	(2/7)
25	QYSGTLSRT	0.178	pTxxR	(6/7)	(2/7)
29	TLSRTPNQI	0.017	pTxxQ	(6/7)	(6/7)
41	EKPSTLSPL	0.262	pTxxP	(7/7)	(7/7)
52	GKKWTEKLA	0.440	pTxxL	(7/7)	(7/7)
77	ISSSTFSFS	0.014	pTxxF	(6/7)	(6/7)
87	KSRVTSSNS	0.568	pTxxN	(4/7)	(6/7)
102	NLMNTPSTV	0.403	pTxxT	(6/7)	(2/7)
105	NTPSTVSTD	0.137	pTxxT	(2/7)	(4/7)
108	STVSTDYLP	0.187	pTxxL	(4/7)	(6/7)
118	HPHRTSSLP	0.464	pTxxL	(7/7)	(7/7)
166	PIQRTSIKK	0.125	pTxxK	(4/7)	(7/7)
178	NASCTLCDE	0.670	pTxxD	(7/7)	(7/7)
213	ISFGTTSKA	0.550	pTxxK	(4/7)	(6/7)
214	SFGTTSKAD	0.945	pTxxA	(6/7)	(4/7)
227	FPFCTKCKK	0.592	pTxxK	(5/7)	(7/7)
233	CKKDTNKAV	0.495	pTxxA	(1/7)	(6/7)
267	ELSITPQSR	0.047	pTxxS	(4/7)	(6/7)
287	GLSYTPVER	0.606	pTxxE	(7/7)	(5/7)
293	VERQTIYSQ	0.935	pTxxS	(6/7)	(6/7)
325	KSNYTFLHS	0.006	pTxxH	(2/7)	(4/7)
347	ILADTSVAL	0.572	pTxxA	(6/7)	(1/7)
371	KDDETKTTL	0.031	pTxxT	(4/7)	(1/7)
373	DETKTTLPL	0.462	pTxxP	(1/7)	(6/7)
374	ETKTTLPLL	0.133	pTxxL	(1/7)	(6/7)
456	LEVFTPIAN	0.098	pTxxA	(4/7)	(1/7)

464	NLRMTTLEA	0.900	pTxxE	(6/7)	(4/7)
465	LRMTTLEAS	0.678	pTxxA	(7/7)	(6/7)
474	VLKCTLNKQ	0.010	pTxxK	(6/7)	(6/7)
498	SDESTTVQK	0.388	pTxxQ	(6/7)	(6/7)
499	DESTTVQKW	0.144	pTxxK	(6/7)	(6/7)
520	EDNITSTLP	0.013	pTxxL	(6/7)	(6/7)
522	NITSTLPIL	0.033	pTxxI	(6/7)	(3/7)
543	GRHETSTFL	0.452	pTxxF	(6/7)	(6/7)
545	HETSTFLGL	0.101	pTxxG	(6/7)	(6/7)
566	HDNDTVIIR	0.022	pTxxI	(6/7)	(6/7)
574	RRGFTLNSG	0.305	pTxxS	(6/7)	(4/7)
585	SRQSTVDSI	0.608	pTxxS	(6/7)	(4/7)
594	QSVLTTISS	0.101	pTxxS	(6/7)	(6/7)
595	SVLTTISSI	0.036	pTxxS	(6/7)	(6/7)
619	QIDFTKLKE	0.036	pTxxK	(4/7)	(6/7)
638	LKALTIKFA	0.439	pTxxF	(4/7)	(6/7)
681	KSSSTQFSP	0.130	pTxxS	(3/7)	(6/7)
691	WLKNTLYPE	0.622	pTxxP	(2/7)	(6/7)
726	YRCFTSFGR	0.744	pTxxG	(4/7)	(6/7)
761	ASSWTFVLE	0.023	pTxxL	(4/7)	(6/7)
766	FVLETLCYS	0.029	pTxxY	(6/7)	(4/7)
790	NDDSTDNEL	0.106	pTxxE	(6/7)	(6/7)
807	DAESTTTIH	0.249	pTxxI	(6/7)	(6/7)
808	AESTTTIHI	0.016	pTxxH	(6/7)	(6/7)
809	**ESTTTIHID**	**0.528**	**pTxxI**	**(6/7)**	**(6/7)**
822	NENATANMV	0.027	pTxxM	(4/7)	(2/7)
833	RNLLTEGEH	0.157	pTxxE	(4/7)	(5/7)
845	ENLETVASS	0.024	pTxxS	(4/7)	(4/7)
869	EEEGTNENE	0.347	pTxxN	(4/7)	(1/7)
894	IDGITRRSS	0.302	pTxxS	(2/7)	(1/7)

✓ Résultats obtenus pour l'interaction entre le domaine FHA1 de Rad53 et Cdc7.

pos	seq	pT proba ?	motif pT ?	pT cons ?	pT+3 cons ?
2	---MTSKTK	0.672	pTxxT	(5/7)	(1/7)
5	MTSKTKNID	0.040	pTxxI	(1/7)	(6/7)
43	IGEGTFSSV	0.518	pTxxS	(7/7)	(7/7)
54	AKDITGKIT	0.265	pTxxI	(6/7)	(6/7)
58	TGKITKKFA	0.299	pTxxF	(7/7)	(6/7)
81	KIYVTSSPQ	0.060	pTxxP	(7/7)	(7/7)
98	LYIMTGSSR	0.068	pTxxS	(7/7)	(7/7)
130	EEFRTFYRD	0.212	pTxxR	(7/7)	(7/7)
167	DIKPTNFLF	0.032	pTxxL	(7/7)	(7/7)
208	NYANTNHDG	0.285	pTxxD	(6/7)	(6/7)
239	SHNQTPPMV	0.550	pTxxM	(6/7)	(7/7)
244	PPMVTIQNG	0.057	pTxxN	(7/7)	(7/7)
262	GVDLTKGYP	0.034	pTxxY	(6/7)	(7/7)
270	PKNETRRIK	0.392	pTxxI	(6/7)	(6/7)
281	NRAGTRGFR	0.339	pTxxF	(7/7)	(7/7)
298	GAQSTKIDI	0.171	pTxxD	(7/7)	(7/7)

334	LELCTIFGW	0.022	pTxxG	(6/7)	(7/7)
382	NKECTIGTF	0.121	pTxxT	(6/7)	(7/7)
385	CTIGTFPEY	0.204	pTxxE	(7/7)	(7/7)
395	VAFETFGFL	0.028	pTxxF	(7/7)	(6/7)
418	PDPKTNMDA	0.215	pTxxD	(3/7)	(6/7)
466	DLLKTPFFN	0.089	pTxxF	(2/7)	(7/7)
476	LNENTYLLD	0.030	pTxxL	(3/7)	(4/7)
484	**DGESTDEDD**	**0.883**	**pTxxD**	**(7/7)**	**(7/7)**

C. Prédiction de la conformation des chaînes latérales.

- ✓ Programmes testés.

Dans ce paragraphe, nous présentons quatre programmes de prédiction de la conformation des chaînes latérales par ordre chronologique d'apparition : SCWRL, FRMSCMFT, SCAP et NCN. Dans la suite, ces quatre programmes seront utilisés pour réaliser une étude des limites actuelles des méthodes de prédiction.

Le programme SCWRL. SCWRL signifie *SideChain Weighted Rotamer Library*. Ce programme, développé au sein de l'équipe de Roland Dunbrack (*Institute for Cancer Research*, Philadelphie, USA), se caractérise dans sa version actuelle par l'utilisation d'une bibliothèque de rotamères dépendante du squelette peptidique, d'une fonction d'énergie très simple et donc rapide, et surtout d'un algorithme astucieux basé sur l'utilisation de la théorie des graphes (Canutescu et al., 2003). L'exécution du programme est particulièrement rapide.

Le programme FRMSCMFT. Développé sous l'impulsion de Joachim Mendès à l'EMBL de Heidelberg, FRMSCMFT est un programme qui considère lors de la phase d'exploration que chaque rotamère peut être flexible autour des positions moyennes que constituent les rotamères principaux (Mendes et al., 1999). On travaille ainsi avec n sous-rotamères par rotamère principal (classiquement n=1000), ce qui augmente considérablement le nombre de configurations possibles. La fonction d'énergie utilisée est très complète et combine la fonction d'énergie du champ de force GROMOS96 avec des données statistiques (Mendes et al., 2001). De par la taille de l'espace des conformations possibles et le choix d'une fonction d'énergie assez complète, l'exécution du programme est lente.

Le programme SCAP. Le programme SCAP est développé au sein de l'université de Columbia (New York, USA) par Zhexin Xiang et Barry Hönig (Jacobson et al., 2002; Xiang and Honig, 2001). Sa bibliothèque est vaste (≈ 7500 rotamères) ; de plus, les longueurs de liaison et les angles de liaison des chaînes latérales ont la possibilité de varier légèrement autour de leurs valeurs théoriques. La fonction d'énergie est simple : elle combine un potentiel statistique traitant des angles de torsion et les contributions Van der Waals. L'algorithme utilisé pour rechercher le minimum énergétique est itératif et basé sur une recherche cyclique.

Le programme NCN. Parmi les quatre programmes présentés, le plus récent est NCN (Peterson et al., 2004). Ce programme, développé dans l'équipe de Joshua Wand au sein de la *Johnson Reasearch Foundation* (Philadelphie, USA), se différencie nettement des autres programmes par la taille de sa bibliothèque de rotamères : près de 50000 rotamères sont référencés, dont plus de 10000 pour l'arginine ! La fonction d'énergie utilisée est une combinaison de paramètres statistiques, comme par exemple la fréquence des rotamères dans la PROTEIN DATA BANK, et de différents potentiels (électrostatique, van der Waals et liaisons hydrogènes). L'algorithme de recherche est basé sur un recuit simulé et permet d'atteindre un minimum local rapidement.

- ✓ Composition de l'ensemble de test.

Les cibles utilisées au départ sont identiques à celles ayant servi à la validation du dernier algorithme SCWRL. Il s'agit d'un ensemble de 180 structures issues de la PROTEIN DATA BANK ayant une seule chaîne, une résolution inférieure ou égale à 1,8 Å et une identité de séquence 2 à 2 ≤ 50%. Toutes ces structures ont été déterminées par diffraction des rayons X.

Les codes PROTEIN DATA BANK sont les suivants : 2ilk, 1bec, 1rb9, 1thv, 1pot, 1tca, 2end, 6cel, 1lam, 8abp, 3ebx, 1thg, 2erl, 1pmi, 1kuh, 1ixh, 1c52, 1a7s, 1gci, 1nkd, 1koe, 1ycc, 1tif, 1orc, 2pth, 3grs, 1bs9, 2por, 1l58, 1dcs, 1tag, 3lzt, 1bd8, 1tyv, 1qnf, 1npk, 1mjr, 1eca, 3pte, 1bfd, 7rsa, 1nls, 1a68, 3cyr, 1lcl, 1uae, 2eng, 1a8d, 2acy, 1xnb, 1vqb, 1msi, 1moq, 1hxn, 3vub, 1hfc, 2hbg, 1aqb, 1kid, 1rzl, 1ctf, 1a8e, 1csh, 1pdo, 1smd, 1aba, 2tgi, 1bm8, 2rn2, 2a0b, 1ako, 2ctc, 1b6g, 1msk, 2qwc, 1bfg, 3nul, 1pda, 1arb, 1wab, 1gof, 1bg6, 2cpl, 1vie, 1cor, 1rhs, 1aie, 1bgf, 3cla, 1ifc, 1vjs, 4xis, 1hal, 1dhn, 1amm, 2sak, 1jer, 1vhn, 1g3p, 1cyo, 1hyp, 1cbn, 3pyp, 1cex, 1ctj, 1a8i, 1mof, 1a6m, 1cnv, 1ryc, 1mrp, 1xjo, 2sn3, 2fdn, 1din, 2bba, 1aru, 1chd, 1cv8, 1bx7, 16pk, 1bdo, 2sns, 3lck, 3seb, 1al3, 1rie, 2dri, 1lbu, 1sbp, 1mml, 1ush, 1edg, 1ads, 2mcm, 1yge, 1whi, 1iab, 1ppn, 1zin, 1lst, 1hka, 1fna, 1poa, 1svy, 3sil, 1fus, 1mun, 1oaa, 1gai, 153l, 119l, 1iuz, 1e70, 1mla, 1bj7, 1ezm, 1ra9, 2igd, 1nox, 1fnc, 1aop, 1opd, 2ayh, 1byi, 1cvl, 1ayl, 1axn, 2cba, 1pgs, 5pti, 1bkf, 1vns, 1aho, 1b6a, 1c3d, 1phb, 1rcf, 1atg, 1cem.

On recherche des protéines ou domaines homologues dont les structures soient connues afin de pouvoir générer des modèles. On souhaite restreindre notre étude aux seules structures

cibles pour lesquelles on retrouve des homologues assez proches et des homologues plus distants, et cela afin de pouvoir générer des modèles plus ou moins proches de la structure native. Pour cela, on effectue un BLASP sur la PROTEIN DATA BANK en ne sélectionnant que les résultats dont la e-value ≤ 0,001. Cette borne a été déterminée de façon à ce qu'on trouve de nombreux homologues potentiels, mais aussi pour que les séquences de la cible et des homologues potentiels s'alignent sur une longueur significative.

Après avoir obtenu la liste des homologues potentiels, les critères pour retenir une structure parmi notre ensemble d'étude sont les suivants :

- la structure retenue doit posséder au moins un homologue avec lequel l'identité de séquence est comprise entre 50 et 70 % ;
- la structure retenue doit posséder au moins un homologue avec lequel l'identité de séquence est comprise entre 30 et 50 % ;
- l'alignement structural et la minimisation s'effectuent correctement (pas d'incident de segmentation pendant l'exécution des programmes).

Parmi les 180 structures de départ, 31 vérifient ces critères. Les résultats pour ces 31 structures cibles sont synthétisés dans le tableau suivant :

Structure cible	Longueur	Homologue 30% à 50% d'identité	Homologue 50% à 70% d'identité.
3ebx	62	1fas	1cod
1iuz	98	1adw	1byp
1rzl	91	1cz2	1mzm
1cyo	88	1fcb	1awp
16pk	415	1hdi	1fw8
7rsa	124	1onc	1rra
1npk	150	2nck	1k44
2ctc	307	1nsa	1dtd
1rcf	169	6nul	1czk
2fdn	55	1ff2	1h98
1nls	237	2lal	1ofs
3lzt	129	1hfx	1jug
3lck	270	1lew	1ksw
1bd8	156	1n0q	1d9s
1gci	269	2prk	1bh6
1ctj	89	1c6s	1cyj
1cbn	46	1bhp	1ed0
2cpl	164	1a58	1h0p
1axn	323	1dm5	1hvd
2sn3	65	1sn4	1cn2

1ycc	108	1jdl	1i54
1bkf	107	1ix5	1r9h
1amm	174	1bd7	1a7h
2cba	258	1koq	1urt
1bfg	126	1g82	1dzd
1poa	118	1mc2	3p2p
1fnc	296	1ogj	1qgz
1xnb	185	1yna	1qh6
1a6m	151	1myt	1lhs
1fus	105	1rtu	1rms
1smd	485	1jd7	1jae

On notera que lorsqu'il existait plusieurs homologues possibles pour une même structure et un même intervalle d'identité de séquence, l'homologue retenu a été sélectionné de façon aléatoire.

Afin de générer des modèles cohérents, les alignements structuraux sont obtenus par le programme DEEP VIEW (Schwede et al., 2003). Grâce à MODELLER 6V2 (Saqi and Sternberg, 1991), on génère 10 modèles à partir de chaque alignement. Pour chaque structure de départ, on dispose donc de 20 modèles construits par homologie. On calcule le *rmsd* entre le squelette de la structure expérimentale et le squelette de chacun des modèles. Les structures pour lesquelles des modèles proches (*rmsd* ≤ 2Å) ont été générés sont : 1a6m, 1axn, 1bfg, 1bkf, 1cbn, 1ctj, 1fnc, 1fus, 1gci, 1poa, 1rcf, 1rzl, 1smd, 1xnb, 1ycc, 2cpl, 2ctc, 2fdn, 2sn3, 3ebx, 3lzt, 7rsa. Les structures pour lesquelles des modèles plus éloignés (2Å ≤ *rmsd* ≤ 4Å) ont été générés sont : 3ebx, 16pk, 1a6m, 1axn, 1bfg, 1ctj, 1cyo, 1fnc, 1fus, 1gci, 1npk, 1smd, 1ycc, 2cba, 2sn3, 3lzt. On procède ensuite à une évaluation succincte des modèles, en regardant la valeur de la fonction objective de MODELLER 6V2 (Saqi and Sternberg, 1991), le score ANOLÉA (Melo and Feytmans, 1998), ainsi que le score donné par FOLDEF (Guerois et al., 2002).

- ✓ Construction des pools de structures

Afin de prédire la conformation des chaînes latérales sans aucun biais, on supprime les chaînes latérales présentes dans les structures cristallographiques et dans les modèles. On dispose ainsi d'un ensemble de squelettes peptidiques sur lesquels les chaînes latérales seront reconstruites. On travaillera dans la suite de cette étude avec 4 pools de structures. Le pool 1 est constitué des structures expérimentales (obtenues par cristallographie) non minimisées et élaguées de leurs chaînes latérales, tandis que le pool 2 est constitué de ces

mêmes structures expérimentales (obtenues par cristallographie) minimisées et élaguées de leurs chaînes latérales. Les pool 3 et le pool 4 sont quant à eux composés de modèles des structures de références construits par homologie. Le pool 3 regroupe des modèles plutôt proches de la structure de référence à laquelle ils se rapportent (0 Å ≤ *rmsd* ≤ 2 Å), minimisés et élagués de leurs chaînes latérales. Pour le pool 4 les modèles considérés sont moins proches (2Å ≤ *rmsd* ≤ 4 Å).

D. Résultats du criblage *in silico* des 5 structures de complexes.

- ✓ Structures de départ.

Les cinq structures initiales ont pour code PDB : 1G6G, 1GXC, 1YJF, 1T15 et 2AZM. Dans chaque cas, seuls les résidus dans une sphère de 15 Å autour de la position criblée sont autorisés à bouger durant la phase d'exploration conformationnelle.

- ✓ Criblage in silico du domaine FHA de Chk2 sur la position pT+3.

La protéine Chk2 est l'orthologue humain de la protéine de levure Rad53. Elle possède également un domaine FHA sur sa portion N-terminale, qui reconnaît spécifiquement les fragments protéiques contenant des motifs pTxxI/L (Li et al., 2002).

En se basant sur la moyenne des scores obtenus par FOLDEF et ROSETTA (**figure 71-d**) on améliore la spécificité de façon significative. Conformément aux résultats du criblage *in vitro* du domaine Chk2, les deux motifs les plus affins sont pTxxI et pTxxL, avec tous deux une variation d'énergie libre estimée à environ -2,0 $kCal.mol^{-1}$. Cependant, l'écart entre ces deux estimations d'énergie libre et celles de cinq autres motifs est inférieur à 0,80 $kCal.mol^{-1}$ (pTxxE, pTxxM, pTxxR, pTxxS, pTxxT) et il a été décrit dans les publications relatives à ROSETTA et FOLDEF que leur erreur moyenne se situait autour de 0,80 $kCal.mol^{-1}$ (Guerois et al., 2002). Ainsi, bien que les prédictions soient correctes dans ce cas, la marge d'erreur associée aux méthodes d'évaluation rend probable la détection de faux positifs.

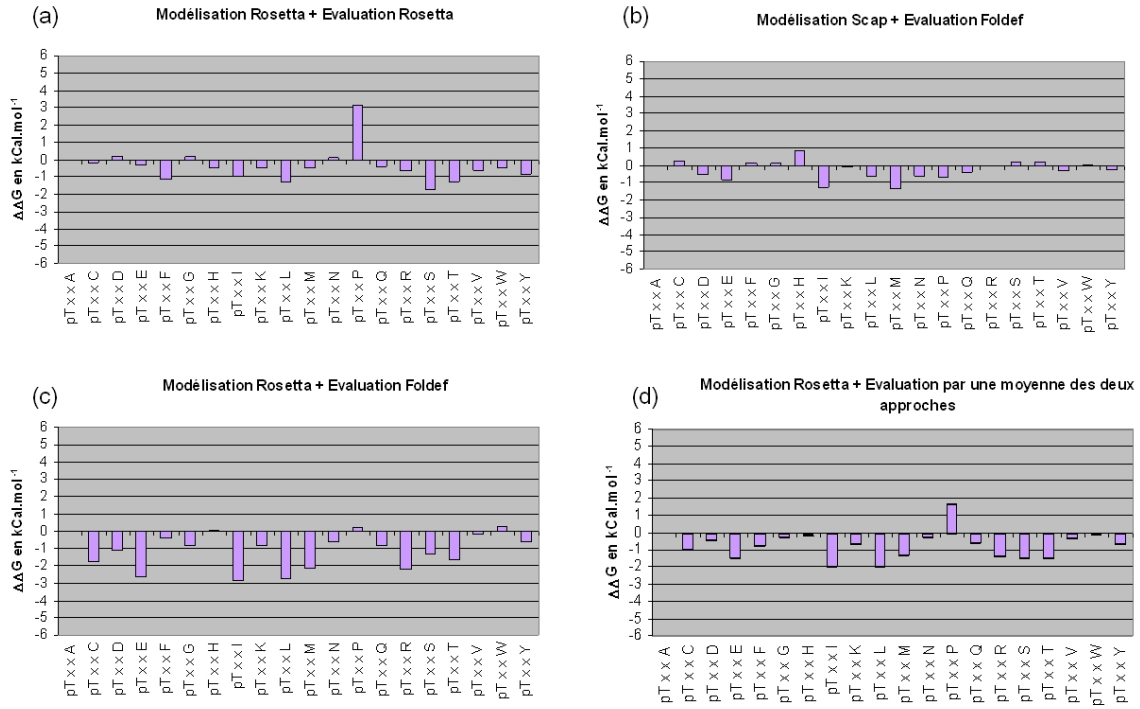


figure 71 : Criblage in silico du domaine FHA N-terminal de la protéine Chk2 sur la position pT+3. Le motif le plus affin pour ce domaine FHA est pTxxI/L. Les graphes **(a) (b) (c)** et **(d)** sont construits de la même façon que dans la figure **figure 63**.

✓ Criblage du domaine FHA de la protéine Pnk sur la position pT-3.

Le domaine FHA de la protéine Pnk se lie *in vivo* à un fragment protéique de Xrcc4 contenant un motif DxxpT (Koch et al., 2004). Des résultats expérimentaux de criblage de bibliothèques de peptides ont confirmé cette capacité du domaine FHA de Pnk à reconnaître spécifiquement une aspartate en position -3 relativement à la thréonine phosphorylée (Bernstein et al., 2005).

En utilisant FOLDEF pour évaluer les structures produites par ROSETTA (**figure 72-b**), des variations d'énergie libre sont mises en évidence, principalement pour les motifs DxxpT ($\Delta\Delta G$ = -1,53 kCal.mol^{-1}) et ExxpT ($\Delta\Delta G$ = -0,92 kCal.mol^{-1}). Cette préférence pour des acides aminés chargés négativement en position pT-3 est concordante avec les résultats expérimentaux. L'utilisation de FOLDEF couplée à SCAP pour l'exploration conformationnelle sélectionne également ces deux résidus acides avec un seuil significatif (**figure 72-c**).

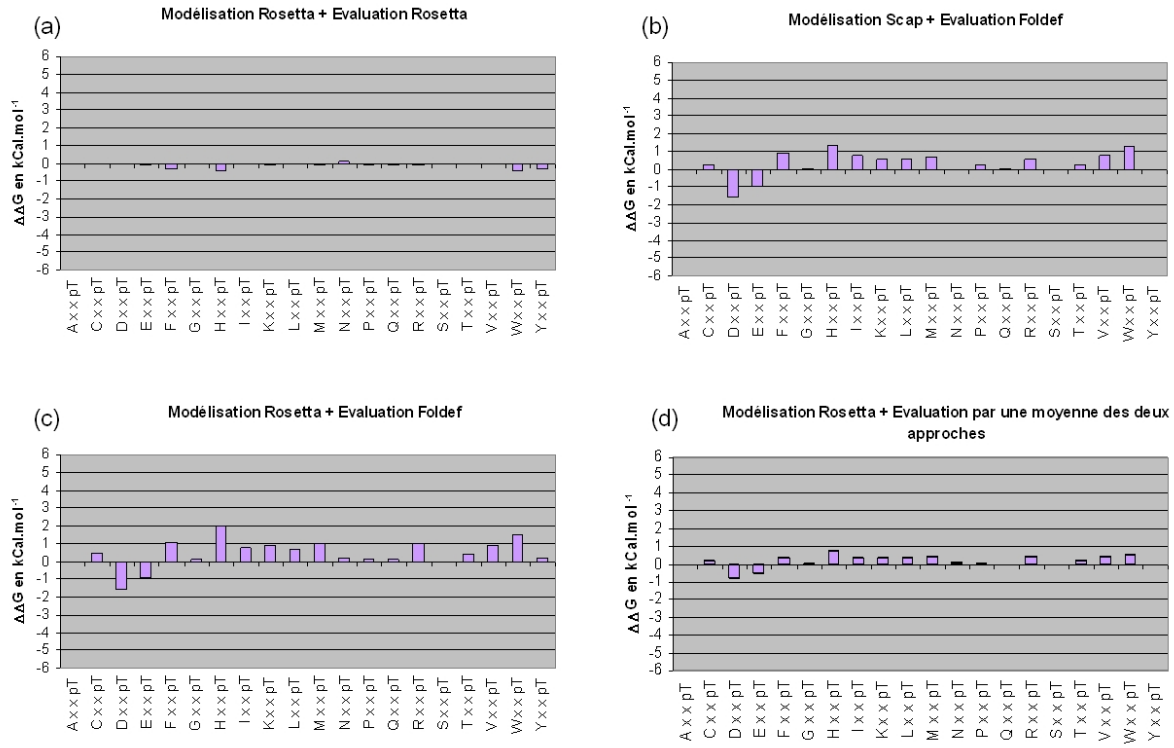


figure 72 : Criblage *in silico* du domaine FHA de la protéine Pnk sur la position pT-3. Le motif le plus affin pour ce domaine FHA est DxxpT. Les graphes **(a) (b) (c)** et **(d)** sont construits de la même façon que dans la figure **figure 63**.

- ✓ Criblage du tandem de domaines BRCT de la protéine Brca1sur la position pS+3.

Le tandem de domaines BRCT de la protéine humaine Brca1 se lie à des motifs contenant une sérine phosphorylée (pS). La présence d'une phénylalanine ou d'une tyrosine en postion pS+3 rend cette interaction particulièrement favorable (Rodriguez et al.,, 2003).

Les prédictions obtenues par criblage *in silico* de ce tandem de domaines BRCT s'avèrent particulièrement fiables. Les quatres stratégies testées précédemment détectent correctement les deux motifs spécifiques (**figure 73-abcd**). On note cependant que l'approche Modélisation ROSETTA +Evaluation ROSETTA surestime l'affinité du motif pSxxW (**figure 73-a**).

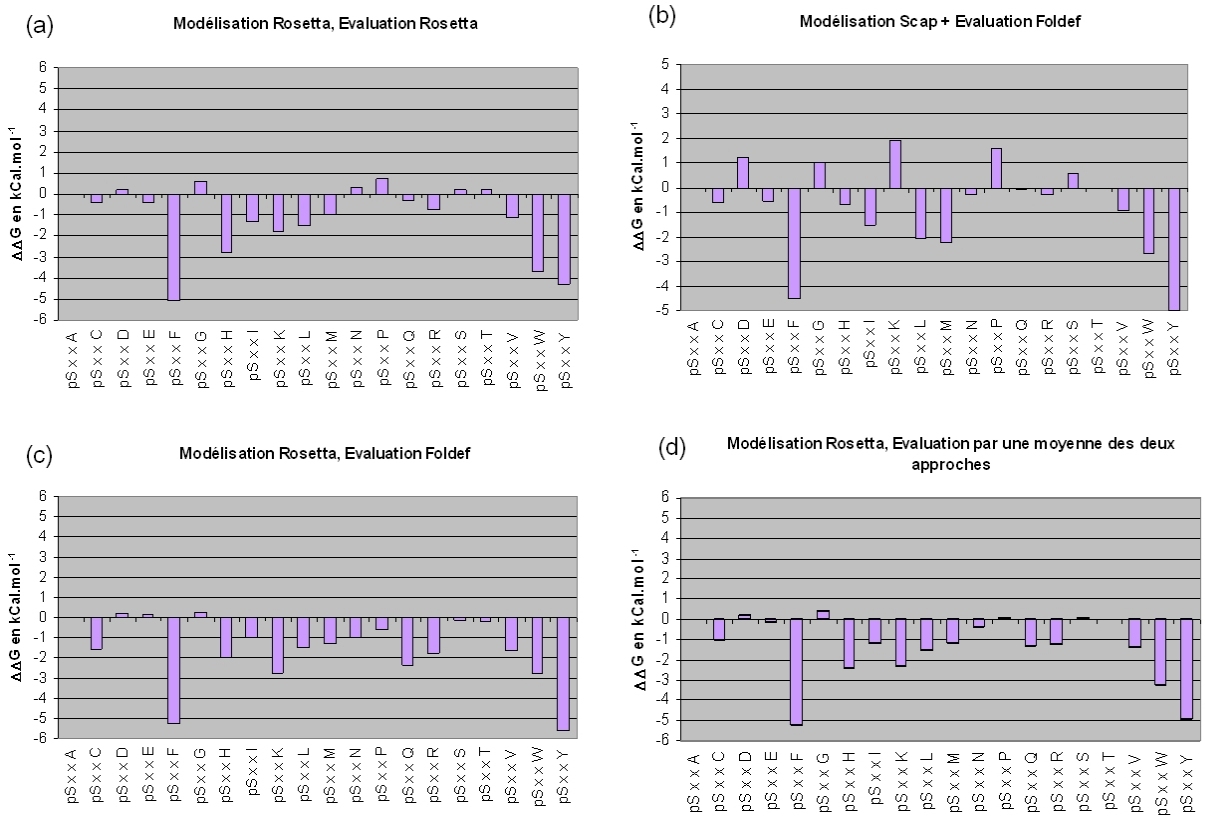


figure 73 : Criblage *in silico* du tandem de domaines BRCT de la protéine Brca1 sur la position pS+3. Le motif le plus affin pour ce module structural est pSxx[Y/F]. Les graphes **(a) (b) (c)** et **(d)** sont construits de la même façon que dans la figure **figure 63**.

- ✓ Criblage du tandem de domaines BRCT de la protéine Mdc1sur la position pS+3.

Il a été établi expérimentalement que le tandem de domaines BRCT de la protéine Mdc1 reconnaît les motifs pSxxY/pSxxF (Lee et al., 2005). Le criblage *in silico* de la position pS+3 sur la structure de ce tandem sélectionne tous les acides aminés aromatiques (phénylalanine, tyrosine, tryptophane), quelle que soit l'approche utilisée (**figure 74-abcd**). Contrairement au résultats obtenus pour le tandem de domaines BRCT de la protéine Brca1, le motif pSxxW est sélectionné quelle que soit la fonction d'énergie utilisée pour l'évaluation des mutants.

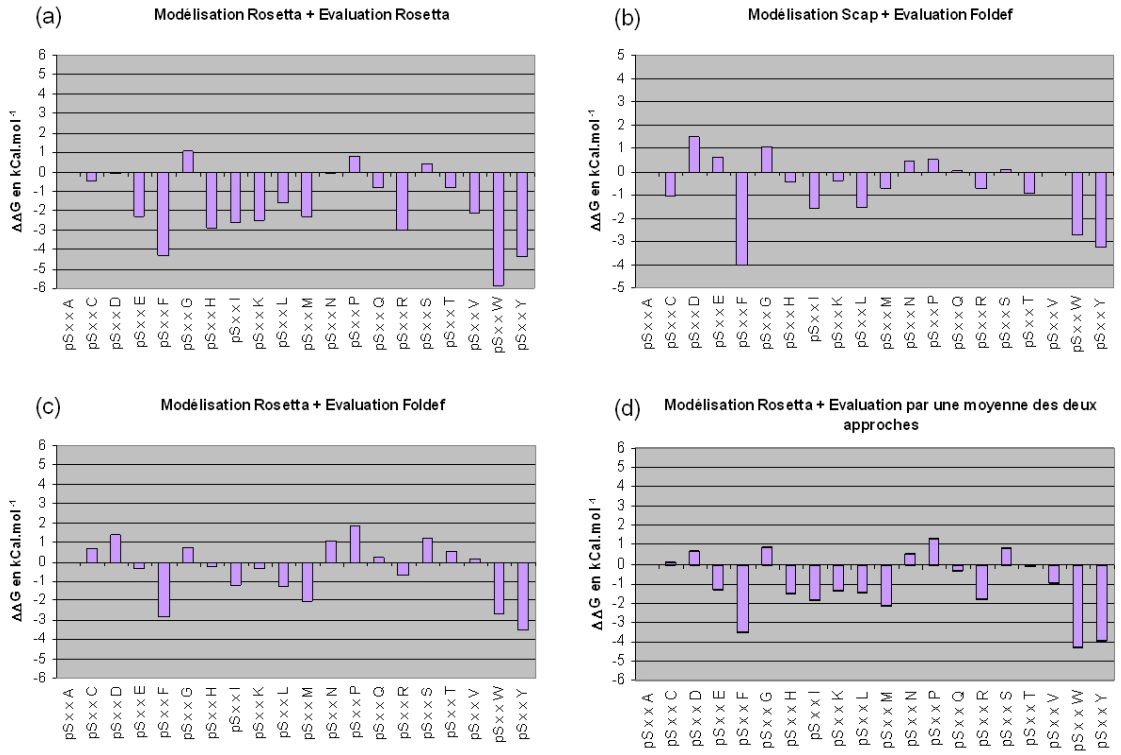


figure 74 : Criblage *in silico* du tandem de domaines BRCT de la protéine Mdc1 sur la position pS+3. Le motif le plus affin pour ce module structural est pSxx[Y/F]. Les graphes **(a) (b) (c)** et **(d)** sont construits de la même façon que dans la figure **figure 63**.

- ✓ Criblage *in silico* du domaine FHA N-terminal de Rad53 sur toute la longueur du peptide.

Les positions {pT-4 ; pT-3 ; pT-2 ; pT+1 ; pT+2 ; pT+3} du peptide en contact avec le domaine FHA1 de Rad53 ont été criblées (**figure 75-abcdef**). On constate que la position pT+3 est celle pour laquelle la sélection d'un acide aminé particulier est la plus favorable : la variation d'énergie libre ΔΔG du motif pTxxD atteint un minimum de -2,65 kCal.mol^{-1}. Ce résultat est cohérent avec les mesures des criblages expérimentaux de bibliothèques de peptides qui ont mis en évidence que la position la plus sélective était la position pT+3 avec une préférence pour les aspartates (Durocher et al., 2000).

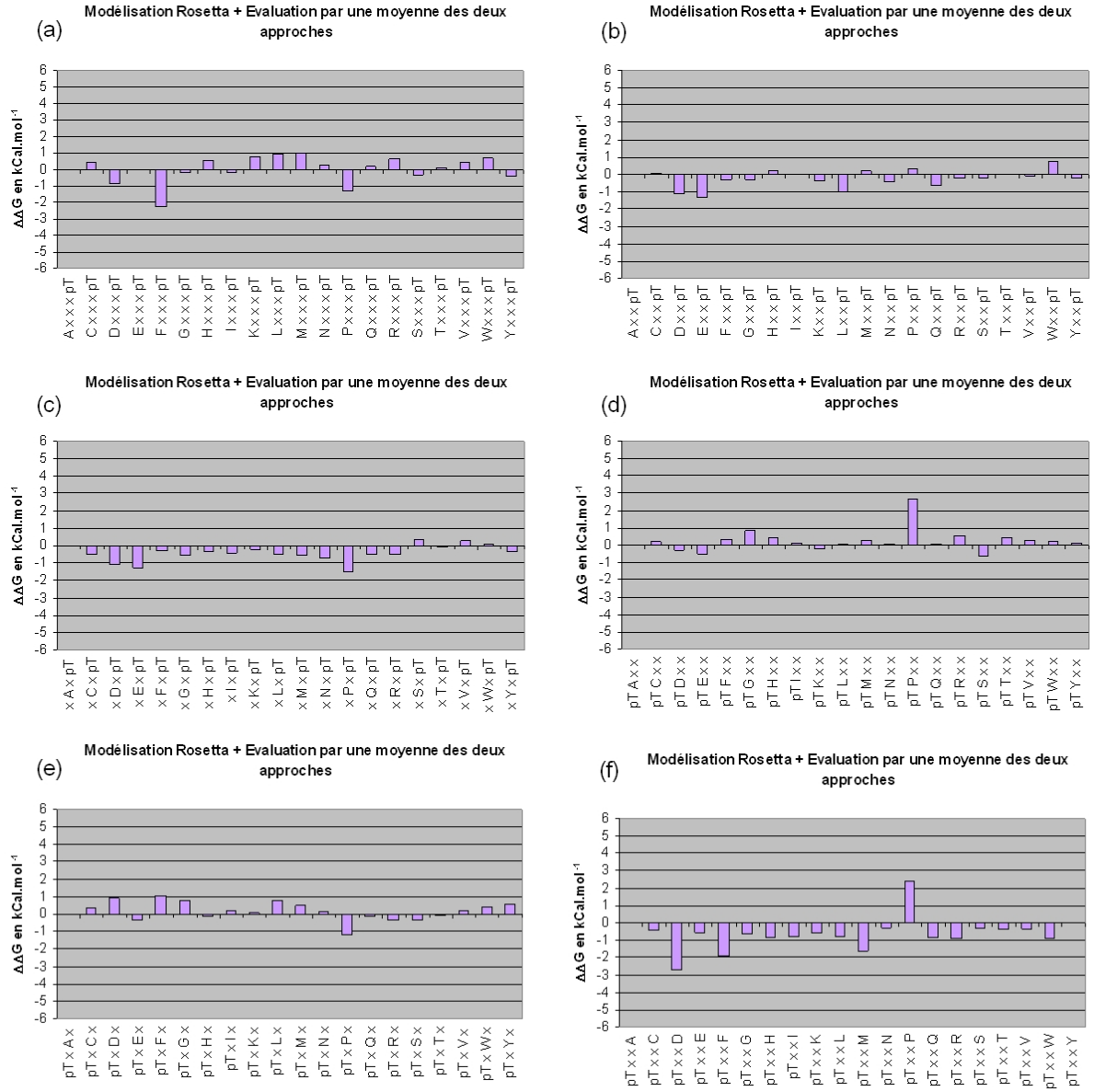


figure 75 : Criblage *in silico* du peptide en contact avec le domaine FHA N-terminal de la protéine Rad53. Les variations d'énergie libre sont estimées par ROSETTA et FOLDEF en kCal.mol^{-1}. Les positions criblées sont **(a)** pT-4, **(b)** pT-3, **(c)** pT-2, **(d)** pT+1, **(e)** pT+2, **(f)** pT+3. La position pT-1, complètement exposée au solvant, n'est pas criblée.

- ✓ Criblage in silico du domaine FHA de Chk2 sur toute la longueur du peptide.

Les positions {pT-3 ; pT-2 ; pT-1 ; pT+1 ; pT+2 ; pT+3} du peptide en contact avec le domaine FHA de Chk2 ont été criblées (**figure 76-abcdef**). Cinq positions montrent une sélectivité particulière : pT-3 pour les aromatiques, pT-1 pour les thréonines, pT+1 pour les tyrosines, pT+2 pour les prolines, et enfin pT+3 pour l'isoleucine et la leucine.

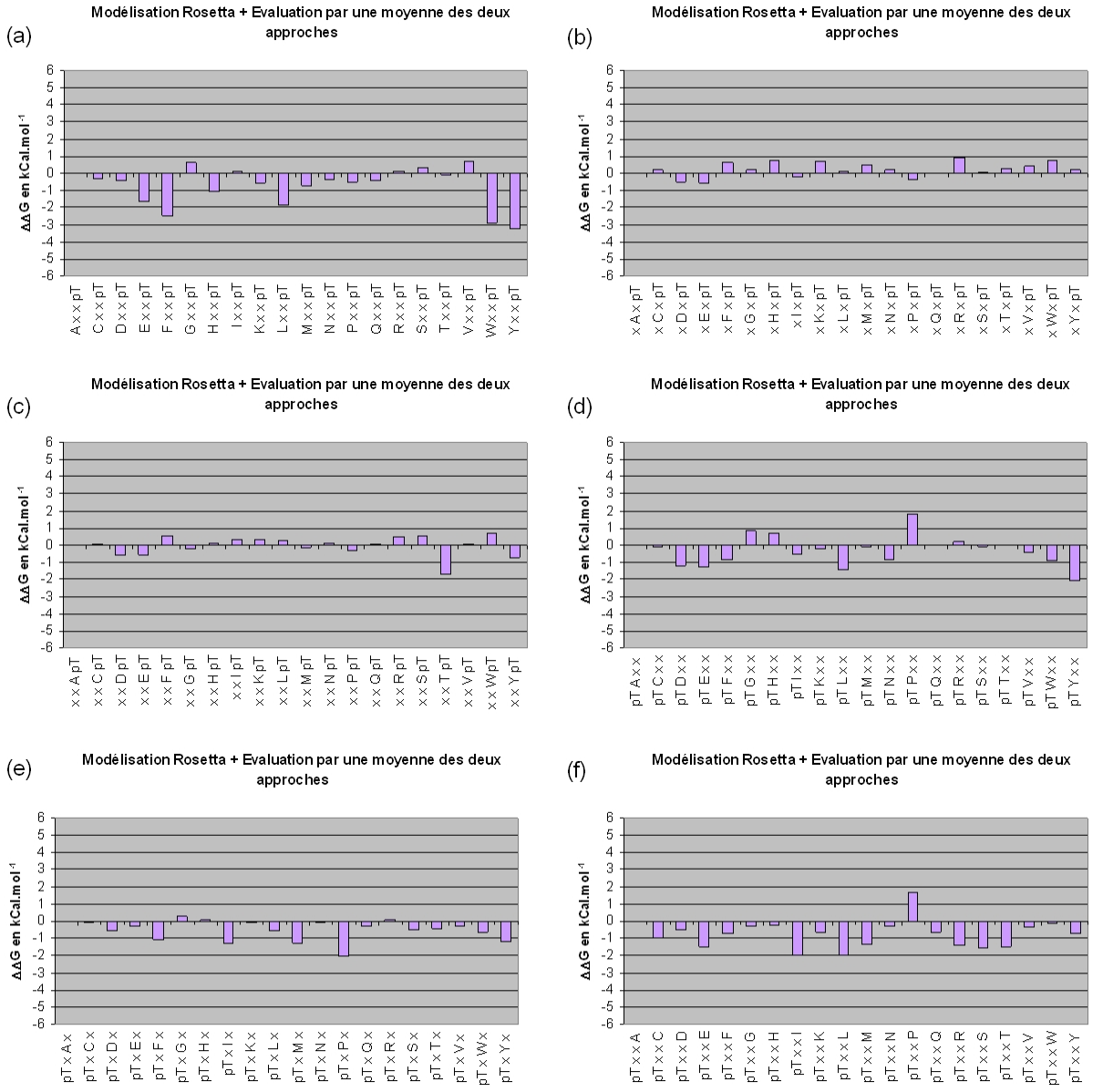


figure 76 : Criblage *in silico* du peptide en contact avec le domaine FHA de la protéine Chk2. Les variations d'énergie libre sont estimées par ROSETTA et FOLDEF en kCal.mol^{-1}. Les positions criblées sont **(a)** pT-3, **(b)** pT-2, **(c)** pT-1, **(d)** pT+1, **(e)** pT+2, **(f)** pT+3.

- ✓ Criblage in silico du domaine FHA de la protéine Pnk sur toute la longueur du peptide.

Les positions {pT-4 ; pT-3 ; pT-2 ; pT-1 ; pT+1 ; pT+2} du peptide en contact avec le domaine FHA de Pnk ont été criblées (**figure 77-abcdef**). Rappelons que ce domaine FHA possède la particularité de reconnaître des aspartates en position pT-3. D'après les résultats du criblage complet *in silico*, les deux positions les plus sélectives sont les positions pT-4 et pT-2.

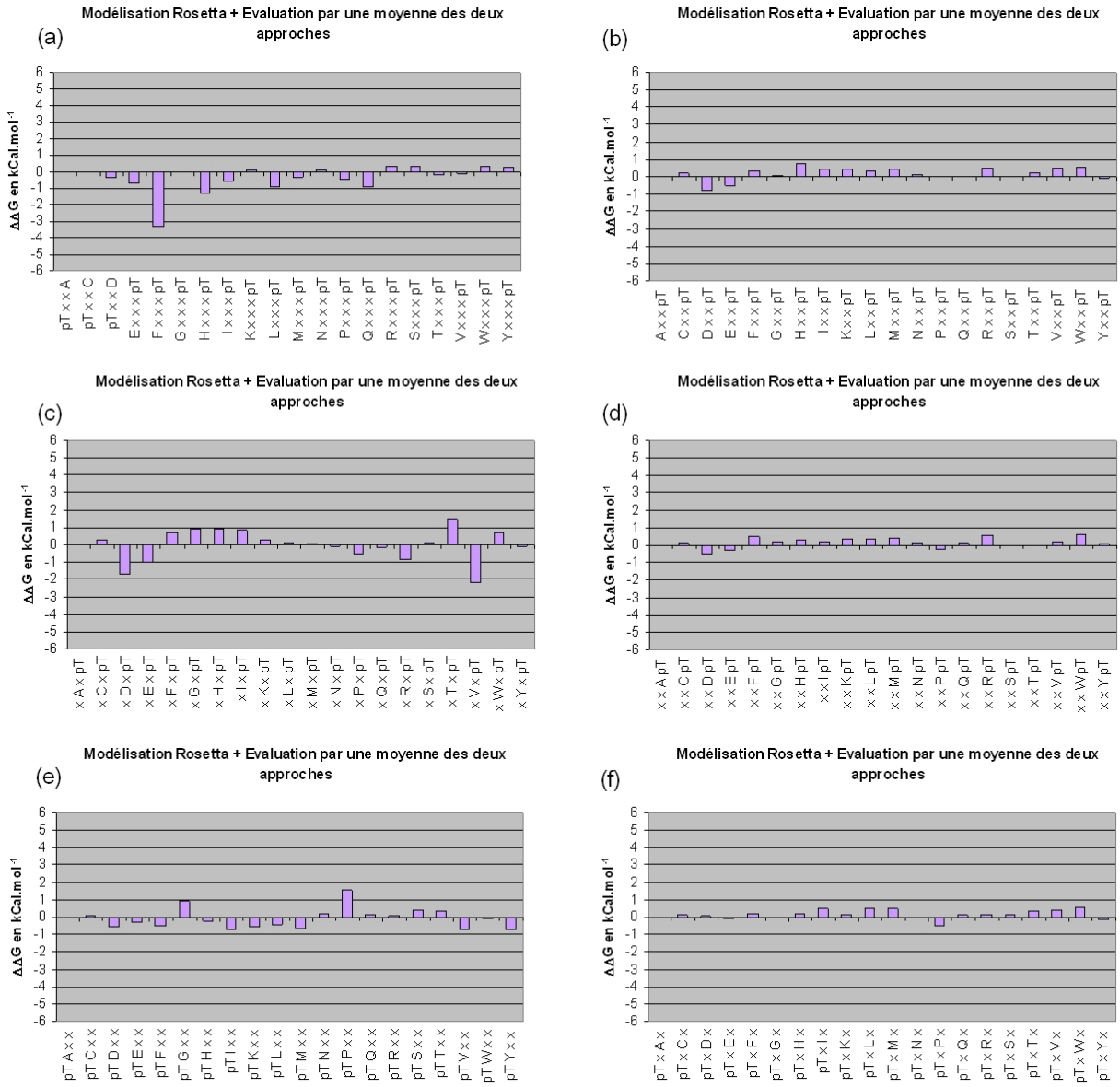


figure 77 : Criblage *in silico* du peptide en contact avec le domaine FHA de la protéine Pnk. Les variations d'énergie libre sont estimées par ROSETTA et FOLDEF en kCal.mol^{-1}. Les positions criblées sont **(a)** pT-4, **(b)** pT-3, **(c)** pT-2, **(d)** pT-1, **(e)** pT+1, **(f)** pT+2.

- ✓ Criblage in silico du tandem de domaines BRCT de la protéine Brca1 sur toute la longueur du peptide.

Les positions {pS-2 ; pS+1 ; pS+2 ; pS+3 ; pS+4} du peptide en contact avec le tandem de domaines BRCT de la protéine Brca1 ont été criblées (**figure 78-abcde**). Les prédictions de ce criblage complet sont de bonne qualité. La position pS+3 est prédite comme la plus sélective, et les motifs reconnus expérimentalement sont effectivement ceux sélectionnés.

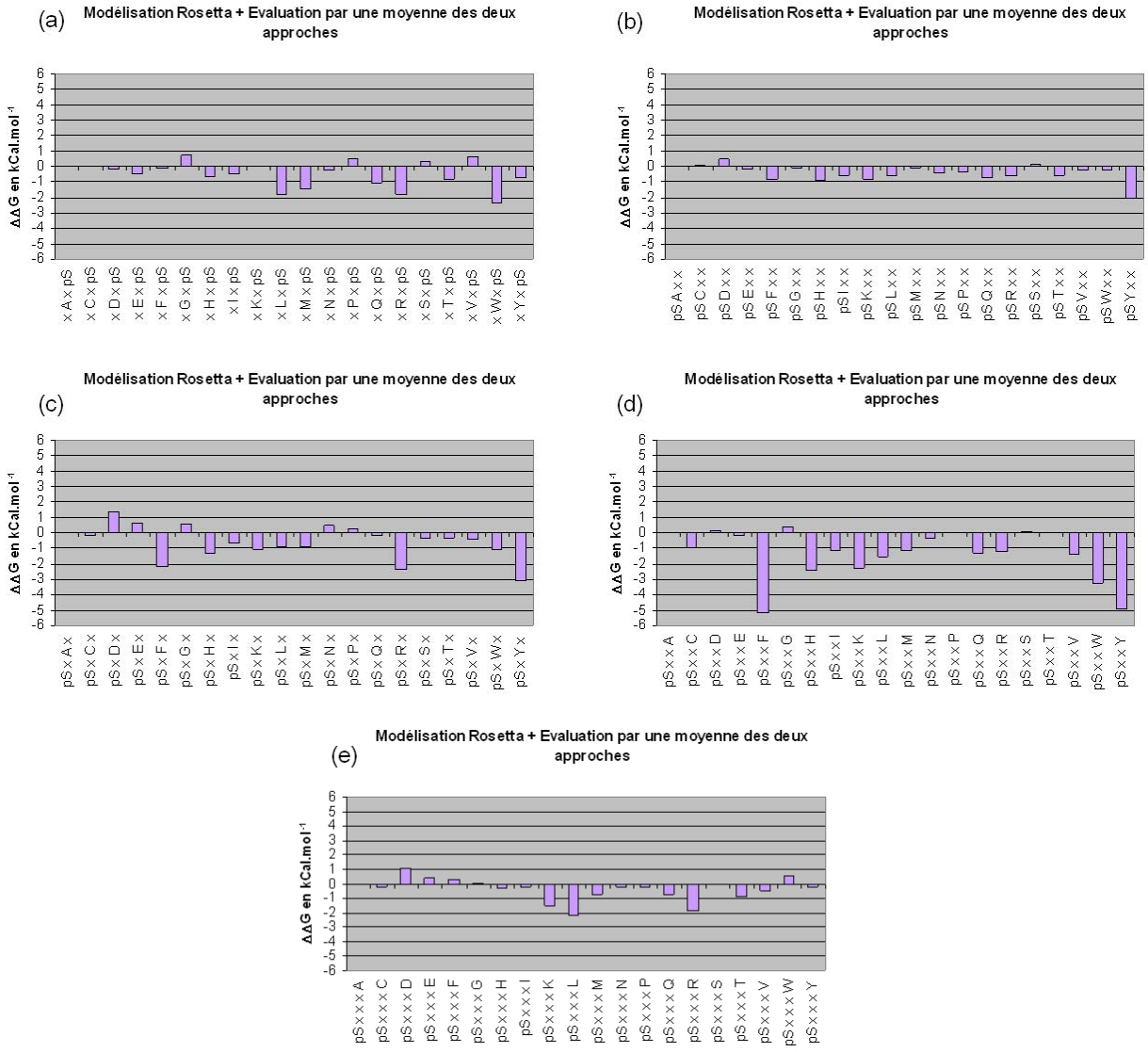


figure 78 : Criblage *in silico* du peptide en contact avec le tandem de domaines BRCT de Brca1. Les variations d'énergie libre sont estimées par ROSETTA et FOLDEF en $kCal.mol^{-1}$. Les positions criblées sont **(a)** pS-2, **(b)** pS+1, **(c)** pS+2, **(d)** pS+3, **(e)** pS+4. La position pS-1, complètement exposée au solvant, n'est pas criblée.

- ✓ Criblage in silico du tandem de domaines BRCT de la protéine Mdc1 sur toute la longueur du peptide.

Pour le tandem de domaines BRCT de la protéine Mdc1, peu de positions ont pu être criblées du fait de l'orientation du peptide. Au final, les prédictions ne concernent que les positions pS+2 et pS+3 (**figure 79-ab**). On constate que la position pS+3 présente une sélectivité forte pour les acides aminés aromatique. Ceci est cohérent avec les résultats expérimentaux qui privilégient les tyrosines et phénylalanines.

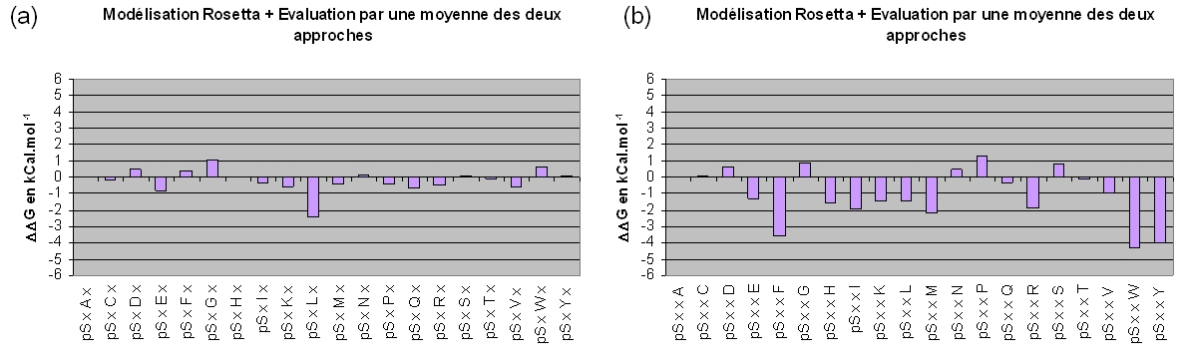


figure 79 : Criblage *in silico* du peptide en contact avec le tandem de domaines BRCT de Mdc1. Les variations d'énergie libre sont estimées par ROSETTA et FOLDEF en $kCal.mol^{-1}$. Les positions criblées sont **(a)** pS+2, **(b)** pS+3. La position pS+1, ainsi que les positions en N-terminal de la sérine phosphorylée, complètement exposées au solvant, ne sont pas criblées.

E. Simulations de dynamique moléculaires.

- ✓ Structure de départ.

La structure du domaine FHA N-terminal de Rad53 non-lié a été étudiée par RMN au sein de l'équipe de Tsai et un ensemble de 20 structures est disponible dans la PROTEIN DATA BANK (code 1K3J). Nous avons sélectionné la structure de plus basse énergie (modèle 1). Cette structure a été superposée à la structure cristallographique du complexe FHA N-terminal de Rad53 – peptide pTxxD. Ceci nous a permis de créer un complexe hybride entre (i) la structure du domaine FHA N-terminal de Rad53 non lié ; et (ii) le peptide issu de la structure

du complexe natif. Sur une région de 15 Å entourant le résidu en position pT+3, le rmsd au niveau du squelette peptidique entre ces deux complexes est de 1,42 Å.

- ✓ Introduction de la thréonine phosphorylée au sein du champs de force OPLS.

La thréonine phosphorylée ne fait pas partie des résidus classiques du champ de force OPLS. Nous l'avons donc ajoutée avec les paramètres suivants, issus de comparaisons avec les paramètres utilisés pour définir les groupements phosphates d'autres molécules dans ce même champ de force.

```
[ TPO ]
 [ atoms ]
     N    opls_238   -0.500     1
     H    opls_241    0.300     1
    CA    opls_224B   0.140     1
    HA    opls_140    0.060     1
    CB    opls_448    0.205     2
    HB    opls_140    0.060     2
   OG1    opls_447   -0.673     2
     P    opls_445    0.808     2
   O1P    opls_446   -0.800     2
   O2P    opls_446   -0.800     2
   O3P    opls_446   -0.800     2
   CG2    opls_135   -0.180     3
  HG21    opls_140    0.060     3
  HG22    opls_140    0.060     3
  HG23    opls_140    0.060     3
     C    opls_235    0.500     4
     O    opls_236   -0.500     4
 [ bonds ]
     N     H
     N    CA
    CA    HA
    CA    CB
    CA     C
    CB    HB
    CB   OG1
    CB   CG2
   OG1     P
     P   O1P
     P   O2P
     P   O3P
   CG2  HG21
   CG2  HG22
   CG2  HG23
     C     O
    -C     N
 [ dihedrals ] ; override some of the typebased dihedrals
     N    CA    CB   OG1    dih_SER_THR_chi1_N_C_C_O
     C    CA    CB   OG1    dih_SER_THR_chi1_CO_C_C_O
```

```
   CA    CB   OG1     P    dih_SER_THR_chi2_C_C_O_P
[ impropers ]
    N    -C    CA     H    improper_Z_N_X_Y
    C    CA    +N     O    improper_O_C_X_Y
```

- ✓ Paramètres des simulations de dynamique moléculaire.

Modélisation sans contraintes ambiguës dans une boîte d'eau. Une simulation de 5ns a été effectuée. La structure du complexe recrée est introduite dans une boîte d'eau rectangulaire de façon à ce que les bords de la boîte soient éloignés de 10 Å des bords de la protéine. La simulation a été effectuée dans un ensemble NVT, la température étant couplée à 300K (algorithme de Berendsen, constante de couplage tT = 0,1ps). L'algorithme PME servi à calculer les interactions électrostatiques (seuil de 0,9 nm). La limite utilisée pour les interactions van der Waals a été fixée à 0,9nm. Le pas d'intégration est de 2fs. Les contraintes d'homologie ont été introduites à l'aide des fonctionnalités développées dans le cadre de la prise en compte des contraintes RMN. Le système de départ comprend au total 17501 atomes. Les atomes de la protéine à plus de 15 Å de la position pT+3 sont gelés durant toute la simulation.

Modélisation avec contraintes ambiguës dans une bulle d'eau. Le système initial est composé du complexe recrée. Les atomes de la protéine à plus de 15 Å de la position pT+3 sont gelés durant la dynamique. Une bulle d'eau de 18 Å de rayon entoure la partie flexible de la protéine (15 Å d'eau libre puis 3 Å d'eau dont les mouvements sont contraints et qui forme une barrière d'étanchéité). Trente simulations de 150 ps chacune ont été lancées. Mis à part la durée de simulation, les paramètres (température, électrostatique, pas d'intégration, etc) sont identiques à ceux utilisés précédemment. Les contraintes ambiguës sont ajoutées et contraintes d'homologie.

- ✓ Algorithmes définissant les contraintes ambiguës

Dans cette partie, nous présentons les algorithmes ayant servi à définir les contraintes ambiguës. Une première partie est consacrée à la définition des différents ensembles. Ensuite est défini l'algorithme de la première phase d'application des contraintes ambiguës, lorsque les contraintes ne sont appliquées qu'entre chaînes latérales. Pour finir, l'algorithme

définissant les contraintes ambiguës de la seconde phase et prenant en compte tous les atomes est présenté.

INITIALISATION DES ENSEMBLES

 # identification des accepteurs

- recherche de l'ensemble A_{sc} de tous les atomes accepteurs asc_1, asc_2… asc_k appartenant à la chaîne latérale d'un acide aminé basique ou polaire dans la bulle d'eau flexible sur laquelle ne s'applique aucune contrainte d'homologie ;
- recherche de l'ensemble A_{bb} de tous les atomes accepteurs abb_1, abb_2, abb_k appartenant au squelette peptidique d'un résidu de la bulle d'eau flexible sur lequel ne s'applique aucune contrainte d'homologie ;

 # identification des donneurs

- recherche de l'ensemble DL_{sc} de tous les atomes lourds donneurs $dlsc_1$, $dlsc_2$… $dlsc_m$ et de l'ensemble DH_{sc} de tous les atomes hydrogènes des donneurs de DL_{sc}, $dhsc_1$, $dhsc_2$… $dhsc_q$ appartenant à la chaîne latérale d'un acide aminé acide ou polaire dans la bulle d'eau flexible sur laquelle ne s'applique aucune contrainte d'homologie ;
- recherche de l'ensemble DL_{bb} de tous les atomes lourds donneurs $dlbb_1$, $dlbb_2$… $dlbb_m$ et de l'ensemble DH_{bb} de tous les atomes hydrogènes des donneurs de DL_{bb}, $dhbb_1$, $dhbb_2$… $dhbb_q$ appartenant au squelette peptidique d'un résidu de la bulle d'eau flexible sur lequel ne s'applique aucune contrainte d'homologie ;

 # identification des hydrophobes

- recherche de l'ensemble H de tous les centres de masse des résidus hydrophobes $h1$, $h2…hp$ de la bulle d'eau flexible ;

ETAPE 1 : {accepteurs chaînes latérales} x {donneurs des chaînes latérales} U {hydrophobes} x {hydrophobes}

 # on cherche à faire un réseau de liaisons hydrogènes en imposant des contraintes entre tous les couples (asc_i,$dlsc_j$) pour i variant entre 1 et m et j variant entre 1 et k (sont exclues les interactions intra-résidu).

Algorithme accepteurs-donneurs-chaines-latérales (A_{sc} , DL_{sc}, DH_{sc}) :

```
        liste_contraintes_ambigues := []
pour i variant de 1 à card(Asc) :
    contrainte_ambigue_l := []
    contrainte_ambigue_h := []
    pour j variant de 1 à card(DLsc) :
        contrainte_ambigue_l.append((asci,dlscj))
    fin pour
    pour j variant de 1 à card(DHsc):
        contrainte_ambigue_h.append((asci,dhscj))
    fin pour
    liste_contraintes_ambigues.append(contrainte_ambigue_l)
    liste_contraintes_ambigues.append(contrainte_ambigue_h)
fin pour
retourner liste_contraintes_ambigues
```

 # on cherche à faire des amas avec les résidus hydrophobes en imposant des contraintes ambiguës entre tous les couples (h_i,h_j) avec i variant de 1 à p et j variant de 1 à p et i !=j.

Algorithme hydrophobes (H) :

```
liste_contraintes_ambigues := []
pour i variant de 1 à p :
    contrainte_ambigue = []
    pour j variant de 1 à card(H) :
        si i != j  alors contrainte_ambigue.append((hi,hj)) fin si
    fin pour
    liste_contraintes_ambigues.append(contrainte_ambigue)
fin pour
retourner liste_contraintes_ambigues
```

ETAPE 2 : {accepteurs} x {donneurs} U {hydrophobes} x {hydrophobes}

 # on cherche à faire un réseau de liaisons hydrogènes en imposant des contraintes entre tous les couples (a_i,dl_j) pour i variant entre 1 et m et j variant entre 1 et k (sont exclues les interactions intra-résidu).

Algorithme accepteurs-donneurs (A_{sc} , A_{bb}, DL_{sc}, DL_{bb}, DH_{sc}, DH_{bb}) :

```
A ← Asc U Abb
DL ← DLsc U DLbb
DH ← DHsc U DHbb

liste_contraintes_ambigues := []
pour i variant de 1 à card(A) :
    contrainte_ambigue_l := []
    contrainte_ambigue_h := []
    pour j variant de 1 à card(DL) :
        contrainte_ambigue_l.append((ai,dlj))
    fin pour
    pour j variant de 1 à card(DH) :
        contrainte_ambigue_h.append((ai,dhj))
    fin pour
    liste_contraintes_ambigues.append(contrainte_ambigue_l)
    liste_contraintes_ambigues.append(contrainte_ambigue_h)
fin pour
retourner liste_contraintes_ambigues
```

 # on cherche à faire des amas avec les résidus hydrophobes en imposant des contraintes ambiguës entre tous les couples (h_i,h_j) avec i variant de 1 à p et j variant de 1 à p et i !=j.

Algorithme hydrophobes (H) :

```
liste_contraintes_ambigues := []
pour i variant de 1 à p :
    contrainte_ambigue = []
    pour j variant de 1 à card(H) :
        si i != j  alors contrainte_ambigue.append((hi,hj)) fin si
    fin pour
    liste_contraintes_ambigues.append(contrainte_ambigue)
fin pour
retourner liste_contraintes_ambigues
```

Publications

Bibliographie

Adkins, M. W., and Tyler, J. K. (2004). The histone chaperone Asf1p mediates global chromatin disassembly in vivo. J Biol Chem 279, 52069-52074.

Agez, M., Chen, J., Guerois, R., van Heijenoort, C., Thuret, J. Y., Mann, C., and Ochsenbein, F. (2007). Structure of the histone chaperone ASF1 bound to the histone H3 C-terminal helix and functional insights. Structure 15, 191-199.

Allen, J. B., Zhou, Z., Siede, W., Friedberg, E. C., and Elledge, S. J. (1994). The SAD1/RAD53 protein kinase controls multiple checkpoints and DNA damage-induced transcription in yeast. Genes Dev 8, 2401-2415.

Alt, J. R., Bouska, A., Fernandez, M. R., Cerny, R. L., Xiao, H., and Eischen, C. M. (2005). Mdm2 binds to Nbs1 at sites of DNA damage and regulates double strand break repair. J Biol Chem 280, 18771-18781.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-3402.

Aparicio, O. M., Stout, A. M., and Bell, S. P. (1999). Differential assembly of Cdc45p and DNA polymerases at early and late origins of DNA replication. Proc Natl Acad Sci U S A 96, 9130-9135.

Argos, P. (1988). An investigation of protein subunit and domain interfaces. Protein Eng 2, 101-113.

Barabasi, A. L., and Albert, R. (1999). Emergence of scaling in random networks. Science 286, 509-512.

Barker, W. C., George, D. G., and Hunt, L. T. (1990). Protein sequence database. Methods Enzymol 183, 31-49.

Bartek, J., Falck, J., and Lukas, J. (2001). CHK2 kinase--a busy messenger. Nat Rev Mol Cell Biol 2, 877-886.

Becker, E., Meyer, V., Madaoui, H., and Guerois, R. (2006). Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response. Bioinformatics 22, 1289-1292.

Bellman, R. (1957). Dynamic Programming (Princeton, New Jersey: Princeton University Press).

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res 28, 235-242.

Bernstein, N. K., Williams, R. S., Rakovszky, M. L., Cui, D., Green, R., Karimi-Busheri, F., Mani, R. S., Galicia, S., Koch, C. A., Cass, C. E., et al. (2005). The molecular architecture of the mammalian DNA repair enzyme, polynucleotide kinase. Mol Cell 17, 657-670.

Bjergbaek, L., Cobb, J. A., Tsai-Pflugfelder, M., and Gasser, S. M. (2005). Mechanistically distinct roles for Sgs1p in checkpoint activation and replication fork maintenance. Embo J 24, 405-417.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294, 1351-1362.

Bousset, K., and Diffley, J. F. (1998). The Cdc7 protein kinase is required for origin firing during S phase. Genes Dev 12, 480-490.

Bradshaw, J. M., Mitaxov, V., and Waksman, G. (2000). Mutational investigation of the specificity determining region of the Src SH2 domain. J Mol Biol 299, 521-535.

Brooijmans, N., Sharp, K. A., and Kuntz, I. D. (2002). Stability of macromolecular complexes. Proteins 48, 645-653.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comp Chem 4, 187-217.

Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjolander, K., and Haussler, D. (1993). Using Dirichlet mixture priors to derive hidden Markov models for protein families. Proc Int Conf Intell Syst Mol Biol 1, 47-55.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5, R6.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., et al. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54, 905-921.

Byeon, I. J., Li, H., Song, H., Gronenborn, A. M., and Tsai, M. D. (2005). Sequential phosphorylation and multisite interactions characterize specific target recognition by the FHA domain of Ki67. Nat Struct Mol Biol 12, 987-993.

Byeon, I. J., Yongkiettrakul, S., and Tsai, M. D. (2001). Solution structure of the yeast Rad53 FHA2 complexed with a phosphothreonine peptide pTXXL: comparison with the structures of FHA2-pYXL and FHA1-pTXXD complexes. J Mol Biol 314, 577-588.

Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L., Jr. (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12, 2001-2014.

Castagnoli, L., Costantini, A., Dall'Armi, C., Gonfloni, S., Montecchi-Palazzi, L., Panni, S., Paoluzi, S., Santonico, E., and Cesareni, G. (2004). Selectivity and promiscuity in the interaction network mediated by protein recognition modules. FEBS Lett 567, 74-79.

Cestra, G., Castagnoli, L., Dente, L., Minenkova, O., Petrelli, A., Migone, N., Hoffmuller, U., Schneider-Mergener, J., and Cesareni, G. (1999). The SH3 domains of endophilin and amphiphysin bind to the proline-rich region of synaptojanin 1 at distinct sites that display an unconventional binding specificity. J Biol Chem 274, 32001-32007.

Chang, M. S., and Benner, S. A. (2004). Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol 341, 617-631.

Chaturvedi, P., Eng, W. K., Zhu, Y., Mattern, M. R., Mishra, R., Hurle, M. R., Zhang, X., Annan, R. S., Lu, Q., Faucette, L. F., et al. (1999). Mammalian Chk2 is a downstream effector of the ATM-dependent DNA damage checkpoint pathway. Oncogene 18, 4047-4054.

Cheadle, C., Ivashchenko, Y., South, V., Searfoss, G. H., French, S., Howk, R., Ricca, G. A., and Jaye, M. (1994). Identification of a Src SH3 domain binding motif by screening a random phage display library. J Biol Chem 269, 24034-24039.

Chivian, D., and Baker, D. (2006). Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 34, e112.

Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., et al. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Res 32, D311-314.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301, 71-76.

Cortes, J., Simeon, T., Ruiz de Angulo, V., Guieysse, D., Remaud-Simeon, M., and Tran, V. (2005). A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 21 Suppl 1, i116-125.

Dahiyat, B. I., and Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. Science 278, 82-87.

Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. J Mol Biol 332, 449-460.

Dayhoff, M. O. (1973). Atlas of Protein Sequence and Structure).

Dayhoff, M. O. (1978). Atlas of Protein Sequence and Structure).

Deane, C. M., Salwinski, L., Xenarios, I., and Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1, 349-356.

Debe, D. A., Danzer, J. F., Goddard, W. A., and Poleksic, A. (2006). STRUCTFAST: protein sequence remote homology detection and alignment using novel dynamic programming and profile-profile scoring. Proteins 64, 960-967.

DePristo, M. A., de Bakker, P. I., Lovell, S. C., and Blundell, T. L. (2003). Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 51, 41-55

Desai-Mehta, A., Cerosaletti, K. M., and Concannon, P. (2001). Distinct functional domains of nibrin mediate Mre11 binding, focus formation, and nuclear localization. Mol Cell Biol 21, 2184-2191.

Desany, B. A., Alcasabas, A. A., Bachant, J. B., and Elledge, S. J. (1998). Recovery from DNA replicational stress is the essential function of the S-phase checkpoint pathway. Genes Dev 12, 2956-2970.

Dhalluin, C., Carlson, J. E., Zeng, L., He, C., Aggarwal, A. K., and Zhou, M. M. (1999). Structure and ligand of a histone acetyltransferase bromodomain. Nature 399, 491-496.

Ding, Z., Lee, G. I., Liang, X., Gallazzi, F., Arunima, A., and Van Doren, S. R. (2005). PhosphoThr peptide binding globally rigidifies much of the FHA domain from Arabidopsis receptor kinase-associated protein phosphatase. Biochemistry 44, 10119-10134.

Dohrmann, P. R., Oshiro, G., Tecklenburg, M., and Sclafani, R. A. (1999). RAD53 regulates DBF4 independently of checkpoint function in Saccharomyces cerevisiae. Genetics 151, 965-977.

Dominguez, C., Boelens, R., and Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125, 1731-1737.

Donaldson, A. D., Fangman, W. L., and Brewer, B. J. (1998). Cdc7 is required throughout the yeast S phase to activate replication origins. Genes Dev 12, 491-501.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., et al. (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics 4, 11.

Dowell, S. J., Romanowski, P., and Diffley, J. F. (1994). Interaction of Dbf4, the Cdc7 protein kinase regulatory subunit, with yeast replication origins in vivo. Science 265, 1243-1246.

Dueber, J. E., Yeh, B. J., Bhattacharyya, R. P., and Lim, W. A. (2004). Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. Curr Opin Struct Biol 14, 690-699.

Dueber, J. E., Yeh, B. J., Chak, K., and Lim, W. A. (2003). Reprogramming control of an allosteric signaling switch through modular recombination. Science 301, 1904-1908.

Dunbrack, R. L., Jr., and Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6, 1661-1681.

Dunbrack, R. L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J Mol Biol 230, 543-574.

Durocher, D., Henckel, J., Fersht, A. R., and Jackson, S. P. (1999). The FHA domain is a modular phosphopeptide recognition motif. Mol Cell 4, 387-394.

Durocher, D., and Jackson, S. P. (2002). The FHA domain. FEBS Lett 513, 58-66.

Durocher, D., Taylor, I. A., Sarbassova, D., Haire, L. F., Westcott, S. L., Jackson, S. P., Smerdon, S. J., and Yaffe, M. B. (2000). The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. Mol Cell 6, 1169-1182.

Eddy, S. R. (1996). Hidden Markov models. Curr Opin Struct Biol 6, 361-365.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics 14, 755-763.

Edgar, R. C., and Sjolander, K. (2004). COACH: profile-profile alignment of protein families using hidden Markov models. Bioinformatics 20, 1309-1318.

Eklund, H., Ingelman, M., Soderberg, B. O., Uhlin, T., Nordlund, P., Nikkola, M., Sonnerstam, U., Joelson, T., and Petratos, K. (1992). Structure of oxidized bacteriophage T4 glutaredoxin (thioredoxin). Refinement of native and mutant proteins. J Mol Biol 228, 596-618.

Emili, A. (1998). MEC1-dependent phosphorylation of Rad9p in response to DNA damage. Mol Cell 2, 183-189.

Emili, A., Schieltz, D. M., Yates, J. R., 3rd, and Hartwell, L. H. (2001). Dynamic interaction of DNA damage checkpoint protein Rad53 with chromatin assembly factor Asf1. Mol Cell 7, 13-20.

Falck, J., Coates, J., and Jackson, S. P. (2005). Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. Nature 434, 605-611.

Feng, S., Chen, J. K., Yu, H., Simon, J. A., and Schreiber, S. L. (1994). Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions. Science 266, 1241-1247.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature 340, 245-246.

Filikov, A. V., Hayes, R. J., Luo, P., Stark, D. M., Chan, C., Kundu, A., and Dahiyat, B. I. (2002). Computational stabilization of human growth hormone. Protein Sci 11, 1452-1461.

Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K. J., Kelley, L. A., MacCallum, R. M., Pawowski, K., et al. (1999). CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins Suppl 3, 209-217.

Fiser, A., Do, R. K., and Sali, A. (2000). Modeling of loops in protein structures. Protein Sci 9, 1753-1773.

Fiser, A., and Sali, A. (2003). ModLoop: automated modeling of loops in protein structures. Bioinformatics 19, 2500-2501.

Fitch, W. M., and Smith, T. F. (1983). Optimal sequence alignments. Proc Natl Acad Sci U S A 80, 1382-1386.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. (2005). Protein interaction mapping: a Drosophila case study. Genome Res 15, 376-384.

Forney, G. D. (1973). The Viterbi Algorithm. Proceedings of the IEEE 61, 268-278.

Gaiser, O. J., Ball, L. J., Schmieder, P., Leitner, D., Strauss, H., Wahl, M., Kuhne, R., Oschkinat, H., and Heinemann, U. (2004). Solution structure, backbone dynamics, and association behavior of the C-terminal BRCT domain from the breast cancer-associated protein BRCA1. Biochemistry 43, 15983-15995.

Garrington, T. P., and Johnson, G. L. (1999). Organization and regulation of mitogen-activated protein kinase signaling pathways. Curr Opin Cell Biol 11, 211-218.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141-147.

Gilbert, D. (2005). Biomolecular interaction network database. Brief Bioinform 6, 194-198.

Ginalski, K., and Rychlewski, L. (2003). Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 31, 3291-3292.

Ginalski, K., von Grotthuss, M., Grishin, N. V., and Rychlewski, L. (2004). Detecting distant homology with Meta-BASIC. Nucleic Acids Res 32, W576-581.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., et al. (2003). A protein interaction map of Drosophila melanogaster. Science 302, 1727-1736.

Glover, J. N., Williams, R. S., and Lee, M. S. (2004). Interactions between BRCT repeats and phosphoproteins: tangled up in two. Trends Biochem Sci 29, 579-585.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science 274, 546, 563-547.

Gowri, V. S., Krishnadev, O., Swamy, C. S., and Srinivasan, N. (2006). MulPSSM: a database of multiple position-specific scoring matrices of protein domain families. Nucleic Acids Res 34, D243-246.

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science 185, 862-864.

Guerois, R., Nielsen, J. E., and Serrano, L. (2002). Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320, 369-387.

Guillemain, G., Ma, E., Mauger, S., Miron, S., Thai, R., Guerois, R., Ochsenbein, F., and Marsolier-Kergoat, M. C. (2007). Mechanisms of checkpoint kinase Rad53 inactivation after a double-strand break in Saccharomyces cerevisiae. Mol Cell Biol.

Guntas, G., Mansell, T. J., Kim, J. R., and Ostermeier, M. (2005). Directed evolution of protein switches and their application to the creation of ligand-binding proteins. Proc Natl Acad Sci U S A 102, 11224-11229.

Guntas, G., Mitchell, S. F., and Ostermeier, M. (2004). A molecular switch created by in vitro recombination of nonhomologous genes. Chem Biol 11, 1483-1487.

Guntas, G., and Ostermeier, M. (2004). Creation of an allosteric enzyme by domain insertion. J Mol Biol 336, 263-273.

Gusfield, D., Balasubramanian, K., and Naor, D. (1992). Parametric optimization of sequence alignment, Paper presented at: Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms (Orlando, Florida, United States: Society for Industrial and Applied Mathematics).

Hantschel, O., Nagar, B., Guettler, S., Kretzschmar, J., Dorey, K., Kuriyan, J., and Superti-Furga, G. (2003). A myristoyl/phosphotyrosine switch regulates c-Abl. Cell 112, 845-857.

Hardy, C. F. (1996). Characterization of an essential Orc2p-associated factor that plays a role in DNA replication. Mol Cell Biol 16, 1832-1841.

Hardy, C. F., and Pautz, A. (1996). A novel role for Cdc5p in DNA replication. Mol Cell Biol 16, 6775-6782.

Henderson, M. J., Munoz, M. A., Saunders, D. N., Clancy, J. L., Russell, A. J., Williams, B., Pappin, D., Khanna, K. K., Jackson, S. P., Sutherland, R. L., and Watts, C. K. (2006). EDD mediates DNA damage-induced activation of CHK2. J Biol Chem 281, 39990-40000.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-10919.

Henikoff, S., and Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. Proteins 17, 49-61.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004). IntAct: an open source molecular interaction database. Nucleic Acids Res 32, D452-455.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180-183.

Hu, F., Alcasabas, A. A., and Elledge, S. J. (2001). Asf1 links Rad53 to control of chromatin assembly. Genes Dev 15, 1061-1066.

Huang, L., and Chiang, D. (2005). Better k-best parsing. Proceedings of the 9th International Workshop on Parsing Technologies (IWPT).

Huang, M., and Elledge, S. J. (1997). Identification of RNR4, encoding a second essential small subunit of ribonucleotide reductase in Saccharomyces cerevisiae. Mol Cell Biol 17, 6105-6113.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci U S A 98, 4569-4574.

Jacobson, M. P., Friesner, R. A., Xiang, Z., and Honig, B. (2002). On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 320, 597-608.

Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 204, 155-164.

Jaroszewski, L., Li, W., and Godzik, A. (2002). In search for more accurate alignments in the twilight zone. Protein Sci 11, 1702-1713.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287, 797-815.

Jones, S., and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. Prog Biophys Mol Biol 63, 31-65.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637.

Kanelis, V., Rotin, D., and Forman-Kay, J. D. (2001). Solution structure of a Nedd4 WW domain-ENaC peptide complex. Nat Struct Biol 8, 407-412.

Kann, M. G., Thiessen, P. A., Panchenko, A. R., Schaffer, A. A., Altschul, S. F., and Bryant, S. H. (2005). A structure-based method for protein sequence alignment. Bioinformatics 21, 1451-1456.

Karplus, K., Barrett, C., and Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. Bioinformatics 14, 846-856.

Karplus, K., and Hu, B. (2001). Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 17, 713-720.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M., and Hughey, R. (2005). SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 61 Suppl 7, 135-142.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299, 499-520.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241-254.

Khanna, K. K., and Jackson, S. P. (2001). DNA double-strand breaks: signaling, repair and the cancer connection. Nat Genet 27, 247-254.

Kihara, M., Nakai, W., Asano, S., Suzuki, A., Kitada, K., Kawasaki, Y., Johnston, L. H., and Sugino, A. (2000). Characterization of the yeast Cdc7p/Dbf4p complex purified from insect cells. Its protein kinase activity is regulated by Rad53p. J Biol Chem 275, 35051-35062.

Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., Zhang, Y., and Bedford, M. T. (2006). Tudor, MBT and chromo domains gauge the degree of lysine methylation. EMBO Rep 7, 397-403.

Kobayashi, J., Antoccia, A., Tauchi, H., Matsuura, S., and Komatsu, K. (2004). NBS1 and its functional role in the DNA damage response. DNA Repair (Amst) 3, 855-861.

Koch, C. A., Agyei, R., Galicia, S., Metalnikov, P., O'Donnell, P., Starostine, A., Weinfeld, M., and Durocher, D. (2004). Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV. Embo J 23, 3874-3885.

Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005). Computational thermostabilization of an enzyme. Science 308, 857-860.

Kortemme, T., Morozov, A. V., and Baker, D. (2003). An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 326, 1239-1259.

Kosiol, C., and Goldman, N. (2005). Different versions of the Dayhoff rate matrix. Mol Biol Evol 22, 193-199.

Kreegipuu, A., Blom, N., and Brunak, S. (1999). PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res 27, 237-239.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637-643.

Krogh, A., Mian, I. S., and Haussler, D. (1994). A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res 22, 4768-4778.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci U S A 97, 10383-10388.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364-1368.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y., and Baker, D. (2001). Conversion of monomeric protein L to an obligate dimer by computational protein design. Proc Natl Acad Sci U S A 98, 10687-10691.

Kuhlman, B., O'Neill, J. W., Kim, D. E., Zhang, K. Y., and Baker, D. (2002). Accurate computer-based design of a new backbone conformation in the second turn of protein L. J Mol Biol 315, 471-477.

Kunin, V., Chan, B., Sitbon, E., Lithwick, G., and Pietrokovski, S. (2001). Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs. J Mol Biol 307, 939-949.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Lautrette, A. (2006) Couplages entre l'assemblage du nucléosome et la réponse cellulaire aux stress génotoxiques, Paris 6.

Lazaridis, T., and Karplus, M. (2000). Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10, 139-145.

Le, S., Davis, C., Konopka, J. B., and Sternglanz, R. (1997). Two new S-phase-specific genes from Saccharomyces cerevisiae. Yeast 13, 1029-1042.

Lee, J. H., and Paull, T. T. (2005). ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. Science 308, 551-554.

Lee, M. S., Edwards, R. A., Thede, G. L., and Glover, J. N. (2005). Structure of the BRCT repeat domain of MDC1 and its specificity for the free COOH-terminal end of the gamma-H2AX histone tail. J Biol Chem 280, 32053-32056.

Lee, S. J., Schwartz, M. F., Duong, J. K., and Stern, D. F. (2003). Rad53 phosphorylation site clusters are important for Rad53 regulation and signaling. Mol Cell Biol 23, 6300-6314.

Leroy, C., Lee, S. E., Vaze, M. B., Ochsenbien, F., Guerois, R., Haber, J. E., and Marsolier-Kergoat, M. C. (2003). PP2C phosphatases Ptc2 and Ptc3 are required for DNA checkpoint inactivation after a double-strand break. Mol Cell 11, 827-835.

Li, H., Byeon, I. J., Ju, Y., and Tsai, M. D. (2004). Structure of human Ki67 FHA domain and its binding to a phosphoprotein fragment from hNIFK reveal unique recognition sites and new views to the structural basis of FHA domain functions. J Mol Biol 335, 371-381.

Li, J., Williams, B. L., Haire, L. F., Goldberg, M., Wilker, E., Durocher, D., Yaffe, M. B., Jackson, S. P., and Smerdon, S. J. (2002). Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. Mol Cell 9, 1045-1054.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., et al. (2004). A map of the interactome network of the metazoan C. elegans. Science 303, 540-543.

Liao, H., Byeon, I. J., and Tsai, M. D. (1999). Structure and function of a new phosphopeptide-binding domain containing the FHA2 of Rad53. J Mol Biol 294, 1041-1049.

Liao, H., Yuan, C., Su, M. I., Yongkiettrakul, S., Qin, D., Li, H., Byeon, I. J., Pei, D., and Tsai, M. D. (2000). Structure of the FHA1 domain of yeast Rad53 and identification of binding sites for both FHA1 and its target protein Rad9. J Mol Biol 304, 941-951.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., et al. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125, 801-814.

Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2003). ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19, 315-316.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. J Mol Biol 285, 2177-2198.

Lockless, S. W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286, 295-299.

Lowndes, N. F., and Toh, G. W. (2005). DNA repair: the importance of phosphorylating histone H2AX. Curr Biol 15, R99-R102.

Luthy, R., Bowie, J. U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature 356, 83-85.

Macdonald, N., Welburn, J. P., Noble, M. E., Nguyen, A., Yaffe, M. B., Clynes, D., Moggs, J. G., Orphanides, G., Thomson, S., Edmunds, J. W., et al. (2005). Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. Mol Cell 20, 199-211.

Mahajan, A., Yuan, C., Pike, B. L., Heierhorst, J., Chang, C. F., and Tsai, M. D. (2005). FHA domain-ligand interactions: importance of integrating chemical and biological approaches. J Am Chem Soc 127, 14572-14573.

Manke, I. A., Lowery, D. M., Nguyen, A., and Yaffe, M. B. (2003). BRCT repeats as phosphopeptide-binding modules involved in protein targeting. Science 302, 636-639.

Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., and Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30, 281-283.

Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. Bioinformatics 17, 359-363.

Matsuoka, S., Huang, M., and Elledge, S. J. (1998). Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. Science 282, 1893-1897.

McGuffin, L. J., and Jones, D. T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19, 874-881.

Mello, J. A., Sillje, H. H., Roche, D. M., Kirschner, D. B., Nigg, E. A., and Almouzni, G. (2002). Human Asf1 and CAF-1 interact and synergize in a repair-coupled nucleosome assembly pathway. EMBO Rep 3, 329-334.

Melo, F., and Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277, 1141-1152.

Mendes, J., Baptista, A. M., Carrondo, M. A., and Soares, C. M. (1999). Improved modeling of side-chains in proteins with rotamer-based methods: a flexible rotamer model. Proteins 37, 530-543.

Mendes, J., Nagarajaram, H. A., Soares, C. M., Blundell, T. L., and Carrondo, M. A. (2001). Incorporating knowledge-based biases into an energy-based side-chain modeling method: application to comparative modeling of protein structure. Biopolymers 59, 72-86.

Meyer, V. (2007) Détection d'homologies lointaines à faibles identités de séquences: Application aux protéines de la signalisation des dommages de l'ADN, Paris 7.

Miller, S., Lesk, A. M., Janin, J., and Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. Nature 328, 834-836.

Miyata, T., Miyazawa, S., and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. J Mol Evol 12, 219-236.

Mizuguchi, K., Deane, C. M., Blundell, T. L., Johnson, M. S., and Overington, J. P. (1998). JOY: protein sequence-structure representation and analysis. Bioinformatics 14, 617-623.

Mizuguchi, K., Deane, C. M., Blundell, T. L., and Overington, J. P. (1998). HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 7, 2469-2471.

Mohana Rao, J. K. (1987). New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int J Pept Protein Res 29, 276-281.

Moult, J. (2005). A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15, 285-289.

Moult, J., Fidelis, K., Rost, B., Hubbard, T., and Tramontano, A. (2005). Critical assessment of methods of protein structure prediction (CASP)--round 6. Proteins 61 Suppl 7, 3-7.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2001). Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5, 2-7.

Moult, J., Fidelis, K., Zemla, A., and Hubbard, T. (2003). Critical assessment of methods of protein structure prediction (CASP)-round V. Proteins 53 Suppl 6, 334-339.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., and Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. Proteins Suppl 1, 2-6.

Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J. T. (1999). Critical assessment of methods of protein structure prediction (CASP): round III. Proteins Suppl 3, 2-6.

Munoz, V., and Serrano, L. (1995). Elucidating the folding problem of helical peptides using empirical parameters. III. Temperature and pH dependence. J Mol Biol 245, 297-308.

Myung, K., Pennaneach, V., Kats, E. S., and Kolodner, R. D. (2003). Saccharomyces cerevisiae chromatin-assembly factors that act during DNA replication function in the maintenance of genome stability. Proc Natl Acad Sci U S A 100, 6640-6645.

Nagar, B., Hantschel, O., Seeliger, M., Davies, J. M., Weis, W. I., Superti-Furga, G., and Kuriyan, J. (2006). Organization of the SH3-SH2 unit in active and inactive forms of the c-Abl tyrosine kinase. Mol Cell 21, 787-798.

Nagar, B., Hantschel, O., Young, M. A., Scheffzek, K., Veach, D., Bornmann, W., Clarkson, B., Superti-Furga, G., and Kuriyan, J. (2003). Structural basis for the autoinhibition of c-Abl tyrosine kinase. Cell 112, 859-871.

Nauli, S., Kuhlman, B., and Baker, D. (2001). Computer-based redesign of a protein folding pathway. Nat Struct Biol 8, 602-605.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48, 443-453.

Nguyen, J. T., Porter, M., Amoui, M., Miller, W. T., Zuckermann, R. N., and Lim, W. A. (2000). Improving SH3 domain ligand selectivity using a non-natural scaffold. Chem Biol 7, 463-473.

Nguyen, J. T., Turck, C. W., Cohen, F. E., Zuckermann, R. N., and Lim, W. A. (1998). Exploiting the basis of proline recognition by SH3 and WW domains: design of N-substituted inhibitors. Science 282, 2088-2092.

Nooren, I. M., and Thornton, J. M. (2003). Diversity of protein-protein interactions. Embo J 22, 3486-3492.

Nooren, I. M., and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein-protein interactions. J Mol Biol 325, 991-1018.

Nozaki, Y., and Bellgard, M. (2005). Statistical evaluation and comparison of a pairwise alignment algorithm that a priori assigns the number of gaps rather than employing gap penalties. Bioinformatics 21, 1421-1428.

Oliver, A. W., Paul, A., Boxall, K. J., Barrie, S. E., Aherne, G. W., Garrett, M. D., Mittnacht, S., and Pearl, L. H. (2006). Trans-activation of the DNA-damage signalling protein kinase Chk2 by T-loop exchange. Embo J 25, 3179-3190.

Overington, J., Johnson, M. S., Sali, A., and Blundell, T. L. (1990). Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc Biol Sci 241, 132-145.

Pawson, T. (2004). Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. Cell 116, 191-203.

Pawson, T., and Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. Science 300, 445-452.

Pellicioli, A., Lee, S. E., Lucca, C., Foiani, M., and Haber, J. E. (2001). Regulation of Saccharomyces Rad53 checkpoint kinase during adaptation from DNA damage-induced G2/M arrest. Mol Cell 7, 293-300.

Peng, T., Zintsmaster, J. S., Namanja, A. T., and Peng, J. W. (2007). Sequence-specific dynamics modulate recognition specificity in WW domains. Nat Struct Mol Biol 14, 325-331.

Peterson, R. W., Dutton, P. L., and Wand, A. J. (2004). Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 13, 735-751.

Petrini, J. H., and Stracker, T. H. (2003). The cellular response to DNA double-strand breaks: defining the sensors and mediators. Trends Cell Biol 13, 458-462.

Pierce, N. A., and Winfree, E. (2002). Protein design is NP-hard. Protein Eng 15, 779-782.

Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. Nucleic Acids Res 24, 3836-3845.

Pike, B. L., Yongkiettrakul, S., Tsai, M. D., and Heierhorst, J. (2004). Mdt1, a novel Rad53 FHA1 domain-interacting protein, modulates DNA damage tolerance and G(2)/M cell cycle progression in Saccharomyces cerevisiae. Mol Cell Biol 24, 2779-2788.

Posern, G., Zheng, J., Knudsen, B. S., Kardinal, C., Muller, K. B., Voss, J., Shishido, T., Cowburn, D., Cheng, G., Wang, B., et al. (1998). Development of highly selective SH3 binding peptides for Crk and CRKL which disrupt Crk-complexes with DOCK180, SoS and C3G. Oncogene 16, 1903-1912.

Prado, F., Cortes-Ledesma, F., and Aguilera, A. (2004). The absence of the yeast chromatin assembly factor Asf1 increases genomic instability and sister chromatid exchange. EMBO Rep 5, 497-502.

Prehoda, K. E., and Lim, W. A. (2002). How signaling proteins integrate multiple inputs: a comparison of N-WASP and Cdk2. Curr Opin Cell Biol 14, 149-154.

Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome: scale-free or geometric? Bioinformatics 20, 3508-3515.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods 24, 218-229.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 Suppl 1, S71-77.

Radley, T. L., Markowska, A. I., Bettinger, B. T., Ha, J. H., and Loh, S. N. (2003). Allosteric switching by mutually exclusive folding of protein domains. J Mol Biol 332, 529-536.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V., et al. (2001). The protein-protein interaction map of Helicobacter pylori. Nature 409, 211-215.

Ramey, C. J., Howar, S., Adkins, M., Linger, J., Spicer, J., and Tyler, J. K. (2004). Activation of the DNA damage checkpoint in yeast lacking the histone chaperone anti-silencing function 1. Mol Cell Biol 24, 10313-10327.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science 297, 1551-1555.

Recht, J., Tsubota, T., Tanny, J. C., Diaz, R. L., Berger, J. M., Zhang, X., Garcia, B. A., Shabanowitz, J., Burlingame, A. L., Hunt, D. F., et al. (2006). Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis. Proc Natl Acad Sci U S A 103, 6988-6993.

Rickles, R. J., Botfield, M. C., Weng, Z., Taylor, J. A., Green, O. M., Brugge, J. S., and Zoller, M. J. (1994). Identification of Src, Fyn, Lyn, PI3K and Abl SH3 domain ligands using phage display libraries. Embo J 13, 5598-5604.

Risler, J. L., Delorme, M. O., Delacroix, H., and Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J Mol Biol 204, 1019-1029.

Rodriguez, M., Yu, X., Chen, J., and Songyang, Z. (2003). Phosphopeptide binding specificities of BRCA1 COOH-terminal (BRCT) domains. J Biol Chem 278, 52914-52918.

Rouse, J., and Jackson, S. P. (2002). Interfaces between the detection, signaling, and repair of DNA damage. Science 297, 547-551.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature 437, 1173-1178.

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature 437, 579-583.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9, 232-241.

Sadreyev, R., and Grishin, N. (2003). COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326, 317-336.

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234, 779-815.

Sallee, N. A., Yeh, B. J., and Lim, W. A. (2007). Engineering modular protein interaction switches by sequence overlap. J Am Chem Soc 129, 4606-4611.

Salwinski, L., and Eisenberg, D. (2003). Computational methods of analysis of protein-protein interactions. Curr Opin Struct Biol 13, 377-382.

Saqi, M. A., and Sternberg, M. J. (1991). A simple method to generate non-trivial alternate alignments of protein sequences. J Mol Biol 219, 727-732.

Schaffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L., and Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15, 1000-1011.

Scholtens, D., and Gentleman, R. (2004). Making sense of high-throughput protein-protein interaction data. Stat Appl Genet Mol Biol 3, Article39.

Scholtens, D., Vidal, M., and Gentleman, R. (2005). Local modeling of global interactome networks. Bioinformatics 21, 3548-3557.

Schwartz, M. F., Duong, J. K., Sun, Z., Morrow, J. S., Pradhan, D., and Stern, D. F. (2002). Rad9 phosphorylation sites couple Rad53 to the Saccharomyces cerevisiae DNA damage checkpoint. Mol Cell 9, 1055-1065.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31, 3381-3385.

Sellers, P. H. (1974). An algorithm for the distance between two finite sequences. J Comb Th A 16, 253-258.

Sellers, P. H. (1974). On the theory and computation of evolutionary distances. SIAM J Appl Math 26, 787-793.

Shannon, C. (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27, 379-423 and 623-656.

Sharp, J. A., Fouts, E. T., Krawitz, D. C., and Kaufman, P. D. (2001). Yeast histone deposition protein Asf1p requires Hir proteins and PCNA for heterochromatic silencing. Curr Biol 11, 463-473.

Shi, J., Blundell, T. L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310, 243-257.

Shroff, R., Arbel-Eden, A., Pilch, D., Ira, G., Bonner, W. M., Petrini, J. H., Haber, J. E., and Lichten, M. (2004). Distribution and dynamics of chromatin modification induced by a defined DNA double-strand break. Curr Biol 14, 1703-1711.

Sidorova, J. M., and Breeden, L. L. (2003). Rad53 checkpoint kinase phosphorylation site preference identified in the Swi6 protein of Saccharomyces cerevisiae. Mol Cell Biol 23, 3405-3416.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34, 82-95.

Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. Proteins 17, 355-362.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., and Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12, 327-345.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-197.

Smith, T. F., Waterman, M. S., and Fitch, W. M. (1981). Comparative biosequence metrics. J Mol Evol 18, 38-46.

Smolka, M. B., Albuquerque, C. P., Chen, S. H., Schmidt, K. H., Wei, X. X., Kolodner, R. D., and Zhou, H. (2005). Dynamic changes in protein-protein interaction and protein phosphorylation probed with amine-reactive isotope tag. Mol Cell Proteomics 4, 1358-1369.

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., and Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. Nature 437, 512-518.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics 21, 951-960.

Soding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33, W244-248.

Sparks, A. B., Quilliam, L. A., Thorn, J. M., Der, C. J., and Kay, B. K. (1994). Identification and characterization of Src SH3 ligands from phage-displayed random peptide libraries. J Biol Chem 269, 23853-23856.

Sparks, A. B., Rider, J. E., Hoffman, N. G., Fowlkes, D. M., Quillam, L. A., and Kay, B. K. (1996). Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. Proc Natl Acad Sci U S A 93, 1540-1544.

Stebbings, L. A., and Mizuguchi, K. (2004). HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. Nucleic Acids Res 32, D203-207.

Stern, D. F., Zheng, P., Beidler, D. R., and Zerillo, C. (1991). Spk1, a new kinase from Saccharomyces cerevisiae, phosphorylates proteins on serine, threonine, and tyrosine. Mol Cell Biol 11, 987-1001.

Stucki, M., Clapperton, J. A., Mohammad, D., Yaffe, M. B., Smerdon, S. J., and Jackson, S. P. (2005). MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. Cell 123, 1213-1226.

Sweeney, F. D., Yang, F., Chi, A., Shabanowitz, J., Hunt, D. F., and Durocher, D. (2005). Saccharomyces cerevisiae Rad9 acts as a Mec1 adaptor to allow Rad53 activation. Curr Biol 15, 1364-1375.

Tagami, H., Ray-Gallet, D., Almouzni, G., and Nakatani, Y. (2004). Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. Cell 116, 51-61.

Toledo, F., and Wahl, G. M. (2006). Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. Nat Rev Cancer 6, 909-923.

Tong, A. H., Drees, B., Nardelli, G., Bader, G. D., Brannetti, B., Castagnoli, L., Evangelista, M., Ferracuti, S., Nelson, B., Paoluzi, S., et al. (2002). A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295, 321-324.

Tramontano, A., and Morea, V. (2003). Assessment of homology-based predictions in CASP5. Proteins 53 Suppl 6, 352-368.

Tramontano, A., and Morea, V. (2003). Exploiting evolutionary relationships for predicting protein structures. Biotechnol Bioeng 84, 756-762.

Tuffery, P., Etchebest, C., and Hazout, S. (1997). Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. Protein Eng 10, 361-372.

Tyler, J. K., Adams, C. R., Chen, S. R., Kobayashi, R., Kamakaka, R. T., and Kadonaga, J. T. (1999). The RCAF complex mediates chromatin assembly during DNA replication and repair. Nature 402, 555-560.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623-627.

van Attikum, H., and Gasser, S. M. (2005). The histone code at DNA breaks: a guide to repair? Nat Rev Mol Cell Biol 6, 757-765.

van den Bosch, M., Bree, R. T., and Lowndes, N. F. (2003). The MRN complex: coordinating and mediating the response to broken chromosomes. EMBO Rep 4, 844-849.

van der Sloot, A. M., Mullally, M. M., Fernandez-Ballester, G., Serrano, L., and Quax, W. J. (2004). Stabilization of TRAIL, an all-beta-sheet multimeric protein, using computational redesign. Protein Eng Des Sel 17, 673-680.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. J Comput Chem 26, 1701-1718.

van Vlijmen, H. W., and Karplus, M. (1997). PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 267, 975-1001.

Vassilev, L. T. (2007). MDM2 inhibitors for cancer therapy. Trends Mol Med 13, 23-31.

Vaze, M. B., Pellicioli, A., Lee, S. E., Ira, G., Liberi, G., Arbel-Eden, A., Foiani, M., and Haber, J. E. (2002). Recovery from checkpoint-mediated arrest after repair of a double-strand break requires Srs2 helicase. Mol Cell 10, 373-385.

Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L., and Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. Nat Struct Biol 9, 485-493.

Verreault, A., Kaufman, P. D., Kobayashi, R., and Stillman, B. (1996). Nucleosome assembly by a complex of CAF-1 and acetylated histones H3/H4. Cell 87, 95-104.

Vetter, S. W., and Zhang, Z. Y. (2002). Probing the phosphopeptide specificities of protein tyrosine phosphatases, SH2 and PTB domains with combinatorial library methods. Curr Protein Pept Sci 3, 365-397.

Vialard, J. E., Gilbert, C. S., Green, C. M., and Lowndes, N. F. (1998). The budding yeast Rad9 checkpoint protein is subjected to Mec1/Tel1-dependent hyperphosphorylation and interacts with Rad53 after DNA damage. Embo J 17, 5679-5688.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans on Information Theory 13, 260-269.

Wallner, B., and Elofsson, A. (2003). Can correct protein models be identified? Protein Sci 12, 1073-1086.

Wang, P., Byeon, I. J., Liao, H., Beebe, K. D., Yongkiettrakul, S., Pei, D., and Tsai, M. D. (2000). II. Structure and specificity of the interaction between the FHA2 domain of Rad53 and phosphotyrosyl peptides. J Mol Biol 302, 927-940.

Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., and Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. Bioinformatics 20, 2138-2139.

Waterman, M. S., and Byers, T. H. (1985). A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. Mathematical biosciences 77, 179-188.

Waterman, M. S., Eggert, M., and Lander, E. (1992). Parametric sequence comparisons. Proc Natl Acad Sci U S A 89, 6090-6093.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature 393, 440-442.

Weng, S., Dong, Q., Balakrishnan, R., Christie, K., Costanzo, M., Dolinski, K., Dwight, S. S., Engel, S., Fisk, D. G., Hong, E., et al. (2003). Saccharomyces Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. Nucleic Acids Res 31, 216-218.

Wiedemann, U., Boisguerin, P., Leben, R., Leitner, D., Krause, G., Moelling, K., Volkmer-Engert, R., and Oschkinat, H. (2004). Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J Mol Biol 343, 703-718.

Williams, B. R., Mirzoeva, O. K., Morgan, W. F., Lin, J., Dunnick, W., and Petrini, J. H. (2002). A murine model of Nijmegen breakage syndrome. Curr Biol 12, 648-653.

Wrabl, J. O., and Grishin, N. V. (2004). Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. Proteins 54, 71-87.

Wysocka, J., Swigut, T., Milne, T. A., Dou, Y., Zhang, X., Burlingame, A. L., Roeder, R. G., Brivanlou, A. H., and Allis, C. D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. Cell 121, 859-872.

Xiang, Z., and Honig, B. (2001). Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 311, 421-430.

Yona, G., and Levitt, M. (2002). Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. J Mol Biol 315, 1257-1275.

Yu, X., Chini, C. C., He, M., Mer, G., and Chen, J. (2003). The BRCT domain is a phospho-protein binding domain. Science 302, 639-642.

Zachariah, M. A., Crooks, G. E., Holbrook, S. R., and Brenner, S. E. (2005). A generalized affine gap model significantly improves protein sequence alignment accuracy. Proteins 58, 329-338.

Zarrinpar, A., Park, S. H., and Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426, 676-680.

Zhang, X., Morera, S., Bates, P. A., Whitehead, P. C., Coffer, A. I., Hainbucher, K., Nash, R. A., Sternberg, M. J., Lindahl, T., and Freemont, P. S. (1998). Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. Embo J 17, 6404-6411.

Zhang, Y., Zhou, J., and Lim, C. U. (2006). The role of NBS1 in DNA double strand break repair, telomere stability, and cell cycle checkpoint control. Cell Res 16, 45-54.

Zhao, X., and Blobel, G. (2005). A SUMO ligase is part of a nuclear multiprotein complex that affects DNA repair and chromosomal organization. Proc Natl Acad Sci U S A 102, 4777-4782.

Zou, L., and Stillman, B. (2000). Assembly of a complex containing Cdc45p, replication protein A, and Mcm2p at replication origins controlled by S-phase cyclin-dependent kinases and Cdc7p-Dbf4p kinase. Mol Cell Biol 20, 3086-3096.

Zuker, M. (1991). Suboptimal sequence alignment in molecular biology. Alignment with error analysis. J Mol Biol 221, 403-420.

ARTICLE 1

Mousson F, Lautrette A, Thuret JY, Agez M, Courbeyrette R, Amigues B, Becker E, Neumann JM, Guerois R, Mann C, Ochsenbein F.
Structural basis for the interaction of Asf1 with histone H3 and its functional implications.
Proc Natl Acad Sci U S A. 2005 Apr 26;102(17):5975-80.

Structural basis for the interaction of Asf1 with histone H3 and its functional implications

Florence Mousson*†, Aurélie Lautrette*, Jean-Yves Thuret‡, Morgane Agez*, Régis Courbeyrette‡, Béatrice Amigues*, Emmanuelle Becker*, Jean-Michel Neumann*, Raphaël Guerois*, Carl Mann‡§¶, and Françoise Ochsenbein*¶

*Service de Biophysique des Fonctions Membranaires and ‡Service de Biochimie et de Génétique Moléculaire, Département de Biologie Joliot-Curie, Commissariat à l'Energie Atomique (CEA/Saclay), F-91191 Gif-sur-Yvette, France

Edited by Gary Felsenfeld, National Institutes of Health, Bethesda, MD, and approved March 17, 2005 (received for review January 7, 2005)

Asf1 is a conserved histone chaperone implicated in nucleosome assembly, transcriptional silencing, and the cellular response to DNA damage. We solved the NMR solution structure of the N-terminal functional domain of the human Asf1a isoform, and we identified by NMR chemical shift mapping a surface of Asf1a that binds the C-terminal helix of histone H3. This binding surface forms a highly conserved hydrophobic groove surrounded by charged residues. Mutations within this binding site decreased the affinity of Asf1a for the histone H3/H4 complex *in vitro*, and the same mutations in the homologous yeast protein led to transcriptional silencing defects, DNA damage sensitivity, and thermosensitive growth. We have thus obtained direct experimental evidence of the mode of binding between a histone and one of its chaperones and genetic data suggesting that this interaction is important in both the DNA damage response and transcriptional silencing.

Asf1 histone chaperone | chromatin | DNA damage | NMR chemical shift mapping | nucleosome assembly

DNA in eukaryotic cells is packaged as nucleosome core particles containing ≈145 bp of DNA wrapped around an octamer comprised of two copies each of histones H2A, H2B, H3, and H4 (1). Assembly of histones into nucleosomes is a tightly orchestrated process (2, 3). Asf1 is a highly conserved histone chaperone that has been linked to both nucleosome assembly and disassembly (4–7). Asf1 interacts with two functional classes of protein: chromatin components, including histone H3 (8), the Hir proteins (9, 10), and the second subunit of CAF-I (5, 11, 12), and checkpoint kinases, including the Rad53 checkpoint kinase in budding yeast (13, 14) and the Tousled-like kinases in metazoans (15). The function of most of these interactions has not been defined. However, a Hir binding region of Asf1 was implicated in telomeric silencing but not required for resistance to genotoxic stress (16). Further work is necessary to determine the functional role of the remaining interactions and, in particular, for defining which Asf1 partners are required for the DNA damage response and for optimal cell growth. In this work, we present the solution structure of the functional N-terminal domain of human Asf1a, and we identify its histone H3 binding site. We show that Asf1 mutants severely defective in histone H3/H4 binding are incompetent in silencing and in providing resistance to DNA damage.

Methods

Protein Production. pETM30 allowed the production of recombinant $(His)_6$-GST-Tev site-fusion proteins in *Escherichia coli* strain BL21 gold (λDE3). Unlabeled and uniformly labeled proteins were obtained as described in ref. 17. After Tev cleavage, the ^{15}N-labeled-H3 (122–135) peptide was further purified by reverse-phase chromatography. The NMR buffer was described in ref. 17. An unlabeled peptide spanning the 122–133 sequence of histone H3 was obtained by chemical synthesis (Epytop, Nîmes, France). The protein concentrations were precisely measured by amino acid analysis.

NMR Structure Determination and Binding Experiments. NMR experiments were carried out on a Bruker DRX-600 spectrometer equipped with a triple-resonance broadband inverse probe or a cryoprobe. ^{1}H, ^{15}N, and ^{13}C resonance assignments were obtained as described in ref. 17. Peak intensities of the ^{15}N- and ^{13}C-edited 3D NOESY-heteronuclear single quantum correlation (HSQC) spectra (with a mixing time of 120 ms) were converted to distance restraints by using ARIA (ambiguous restraints for iterative assignment) (18, 19). One hundred thirteen (ϕ,ψ) restraints were derived by using TALOS (20) and 57 ϕ restraints by using $^3J_{HNH\alpha}$ couplings constants derived from the 3D HNHA spectrum (21); on the basis of the empirical Karplus relation, they were set to $-60 \pm 40°$ for $J_{HN\text{-}Ha} < 6$ Hz and to $-120 \pm 60°$ for $J_{HN\text{-}Ha} > 8$ Hz. An overview of the constraints is given in Table 1, which is published as supporting information on the PNAS web site. ARIA was used to calculate 20 structures, starting from random conformations with the standard procedure. Cumulative chemical shift variation upon binding was calculated as $\Delta\delta = [(\delta_{HN}{}^b - \delta_{HN}{}^f)^2 + (2.75(\delta_{H\alpha}{}^b - \delta_{H\alpha}{}^f))^2 + (0.17(\delta_N{}^b - \delta_N{}^f))^2]^{1/2}$, where b and f refer to the bound and free form, respectively. The scaling factors normalize the magnitude of the ^1H_N, $^1H\alpha$, and ^{15}N chemical shift changes (in ppm) (22). These factors were established from estimates of atom-specific chemical shift ranges in proteins: 5.5 ppm for ^1H_N, 2 ppm for $^1H\alpha$, and 32 ppm for ^{15}N. Assignments of the HSQC spectra of complexed Asf1 (1–156) and H3 (122–135) were obtained with 3D ^{15}N edited NOESY-HSQC and total correlation spectroscopy-HSQC spectra.

Phenotypic Testing of *asf1* Mutants. Wild-type and *asf1* mutants containing a 13myc C-terminal tag were expressed from the endogenous *ASF1* promoter on the centromeric *TRP1* vector pRS314. These plasmids were introduced into the following four yeast strains to test their ability to complement the DNA damage sensitivity, the thermosensitive growth defect, and the transcriptional silencing defect of *asf1*Δ or *asf1*Δ *cac2*Δ mutants. UCC6562 (a generous gift of Dan Gottschling, Fred Hutchinson Cancer Research Center, Seattle) = *Matα ade2*Δ*::hisG lys2Δ0 met15Δ0 trp1-Δ63 his3Δ200 ura3-52 leu2Δ0 DIA5-1 ppr1::LYS2 adh4::TEL(VIIL)-URA3 asf1::HIS3*; CMY1317 = as for

BIOCHEMISTRY

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: TAP, tandem affinity purification.

Data deposition: The atomic coordinates and structure factors for human Asf1a have been deposited in the Protein Data Bank, www.pdb.org (PDB ID code 1TEY). The NMR chemical shifts for human Asf1a have been deposited in the BioMagResBank, www.bmrb.wisc.edu (accession no. 6298).

†Present address: Department of Physiological Chemistry, Division of Biomedical Genetics, University Medical Center Utrecht, Universiteitsweg 100, 3584CG Utrecht, The Netherlands.

§Present address: Department of Biochemistry and Molecular Biology, F. Edward Hébert School of Medicine, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814-4799.

¶To whom correspondence may be addressed. E-mail: cmann@usuhs.mil or ochsenbe@dsvidf.cea.fr.

UCC6562, but *asf1*::*HIS3 cac2*::*kanMX*; CMY1312 = *Matα ade2-101 lys2-801 trp1-Δ63 his3Δ200 ura3-52 leu2Δ1 hmlα*::*URA3 asf1*::*HIS3 cac2*::*kanMX;* CMY1314 = *Matα ade2-101 lys2-801 trp1-Δ63 his3Δ200 ura3-52 leu2Δ1 hmrα*::*URA3 asf1*::*HIS3 cac2*::*kanMX.* Preparation of protein extracts, SDS/PAGE, and immunoblotting analysis were performed as described in ref. 23. Silencing of the *URA3* reporter gene was monitored by the ability of strains to grow on a medium containing 5-fluoroorotic acid as described in ref. 24.

GST Pull-Down Assays. Twenty micrograms of purified $(His)_6$-GST-fusion proteins were immobilized on reduced glutathione agarose beads and equilibrated with 200 μl of buffer H150 (20 mM Hepes-NaOH, pH 7.4/150 mM NaCl/0.5% Nonidet P-40/1 mM EDTA). One microgram of purified native chicken histones H3-H4, kindly provided by A. Prunell (Institut Jacques Monod, Paris) (25), was added to beads. Beads were washed successively with buffers identical to buffer H150 with increasing NaCl concentration up to 2 M. Bound H3 proteins were analyzed by SDS/PAGE and revealed by an antibody against the carboxyl terminus of histone H3 (Abcam, Cambridge, U.K.). $(His)_6$-GST fusion proteins were revealed by a polyclonal antibody against the $(His)_6$ tag (Novagen).

Tandem Affinity Purification (TAP) Tag Purifications of Asf1 and Asf1-V94R. pRS304-*TRP1-ASF1–13myc* and pRS304-*TRP1-asf1-V94R-13myc* were linearized with MluI and integrated at their normal chromosomal locus by transforming strain W303–1b *asf1*::*kanMX* and selecting Trp+ colonies. The 13myc tag was replaced with the TAP tag by transforming these strains with an *ASF1-TAP*::*HIS3MX6* PCR cassette. This cassette contained 111 bp of *ASF1* coding sequence upstream of the stop codon in front of the TAP sequence and 110 bp of *ASF1* 3′ noncoding sequences downstream of the *HIS3MX*6 sequence. Two liters of wild-type and *asf1-V94R* cells were grown to an optical density of 2 in a rich yeast extract/peptone/dextrose medium and the TAP-tagged proteins were purified by the standard protocol in ref. 26 after breaking cells in an Eaton press. Proteins copurifying with Asf1 were separated by SDS/PAGE and stained with Coomassie blue. The bands corresponding to Asf1 and the histones H3 and H4 were identified by a combination of mass spectrometry and immunoblotting (data not shown).

Supporting Information. A description of the DNA constructs and methods of site-specific mutagenesis used in this work is provided as *Supporting Methods*, which is published as supporting information on the PNAS web site.

Results

The Structure of the Asf1 N-Terminal Domain Is Highly Conserved. We determined the structure of the conserved N-terminal domain of human Asf1a (amino acid 1–156) by using multidimensional NMR spectroscopy (Fig. 1*a* and Table 1). The structure comprises 10 β-strands organized into an Ig-like fold composed of three β-sheets topped by two short α-helices. The ^{15}N relaxation parameters R_1, R_2, and the $^{15}N\{^1H\}$-nuclear Overhauser effect are consistent with a monomeric globular domain that presents limited internal motions with significant flexibility in the loops connecting β3-β4, β4-β5, and β8-β9 (see Fig. 6, which is published as supporting information on the PNAS web site). In contrast, the C-terminal region of human Asf1a is fully unfolded (see Fig. 7, which is published as supporting information on the PNAS web site) as indicated by the null or negative values of $^{15}N\{^1H\}$-nuclear Overhauser effects (data not shown).

The x-ray crystal structure of the N-terminal domain of *S. cerevisiae* Asf1 was recently described in ref. 16. Although the sequence of the Asf1 N-terminal domain is highly conserved (58% sequence identity between *S. cerevisiae* Asf1 and human Asf1a), some genetic data highlight functional differences. The human protein does not complement *asf1*Δ mutants of either *S. pombe* (27) *or S. cerevisiae* (data not shown), whereas the *S. cerevisiae* Asf1 N-terminal domain does complement both mutants (16, 27). The NMR solution structure of the N-terminal domain of human Asf1a superimposes well on the x-ray crystal structure of the corresponding domain from *S. cerevisiae* with a root mean square deviation of 1.78 Å for all Cα atoms and 0.83 Å for the Cα atoms of the 10 β-strands (Fig. 1*b*). The major dissimilarities are observed for the A_{48}-Y_{53} and N_{80}-G_{91} loops connecting strands β4-β5 and strands β6-β7, respectively, and the sequences of these two loops are among the most divergent regions between the human and yeast proteins (only 24% identity) (Fig. 8, which is published as supporting information on the PNAS web site). These variations may help in identifying the important factors responsible for the functional differences between the human and yeast proteins.

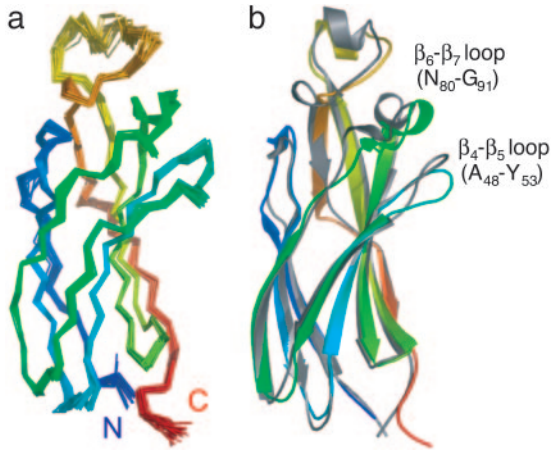


Fig. 1. The structure of the human Asf1a N-terminal domain is well conserved. (*a*) Bundle of 20 structures of human Asf1a (1–156) calculated as indicated in *Methods*. The coloring scheme indicates the relative position of the backbone atoms with the warmest colors being closest to the carboxyl terminus. (*b*) Schematic ribbon diagram (generated with PYMOL, DeLano Scientific, South San Francisco, CA) of the human Asf1a (1–156) structure closest to the mean (colored as in *a*) superimposed on the crystallographic structure of the homologous domain of *S. cerevisiae* Asf1 (16) (gray).

Interaction of Asf1 with the C-Terminal Helix of Histone H3. The histone H3/H4 complex is the best-characterized and most highly conserved partner of Asf1. However, nothing is known concerning the structural basis of this interaction or the functional pathways for which it is required. The C-terminal amino acids 97–135 of histone H3 were implicated in this interaction in a two-hybrid screen (8). This segment spans half of the helix α2 and full-length helix α3 of histone H3 in the crystal structure of the nucleosome (1). We characterized the interaction between human Asf1a and histone H3 by NMR. Addition of the H3 peptide spanning residues 97–135 to ^{15}N-Asf1a (1–156) led to line broadening due to an intermediate exchange rate, preventing further NMR characterization (data not shown). Further two-hybrid analyses indicated that shorter H3 C-terminal peptides could still interact with Asf1a (data not shown). We thus tried the shortened H3 C-terminal peptide (amino acid 122–135) in our NMR experiments and found that it led to significant variations in the Asf1 ^{15}N-HSQC spectrum with a rapid exchange rate (Fig. 2*a*). Transposing the mean square chemical shift variation upon titration (Fig. 2*b*) on the protein structure revealed a well defined binding surface on Asf1 located in a concave groove of the protein (Fig. 3). This groove is highly conserved, principally hydrophobic with residue V94 at its

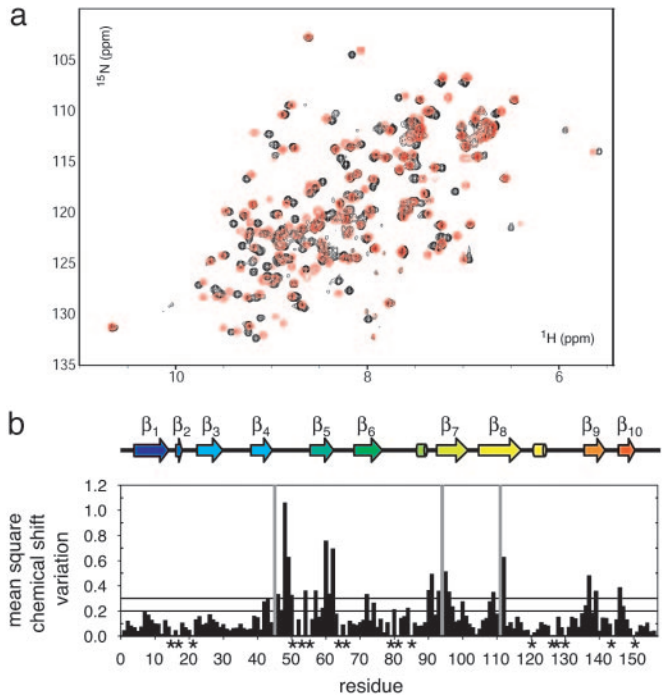


Fig. 2. Interaction of human Asf1a (1–156) with histone H3 (122–135). (*a*) Superposition of the HSQC spectrum of human Asf1a (1–156) (80 μM) with (red) and without (black) an excess of histone H3 (122–135) (160 μM) at 298 K. (*b*) Mean-square chemical shift variation calculated as described in *Methods* for human Asf1a (1–156) upon binding of H3 (122–135) as a function of the sequence. Horizontal lines at 0.2 and 0.3 show the cutoff for the color-coded representation of the protein surface in Fig. 3. The chemical exchange rate was rapid for most of the residues except the highly affected residues 45, 94, and 111, indicated with a vertical gray bar, that exhibited an intermediate exchange rate and severe line broadening due to chemical exchange that led to their absence from the ^{15}N-HSQC spectrum of the complex. Stars indicate missing values (proline residues or unassigned because of signal overlaps).

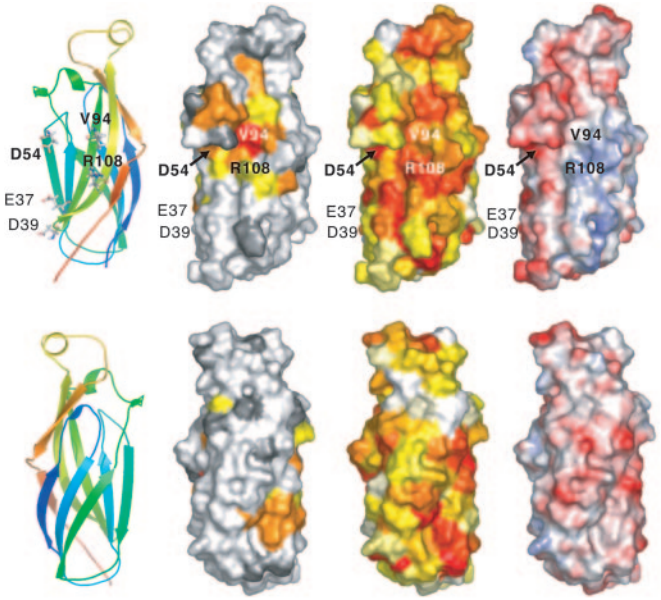


Fig. 3. Structure mapping of the chemical shift variation of Asf1a upon H3 (122–135) binding. Two orientations of the protein are presented in *Upper* and *Lower*. (*Left*) Ribbon representation of the protein. (*Left Center*) Surface color-coded representation of the mean-square chemical shift variation of Asf1a upon H3 (122–135) binding. The three line-broadened residues are shown in red, residues with $\Delta\delta > 0.3$ in orange, residues with $0.2 < \Delta\delta < 0.3$ in yellow, residues with $\Delta\delta < 0.2$ in white and undetermined residues in gray. (*Right Center*) Surface color-coded representation of the residue conservation of Asf1. The conservation was determined by using the RATE4SITE program (38) based on an alignment of 17 Asf1 sequences. The color code is a gradient from red (fully conserved residues) to white (unconserved residues). (*Right*) Color-coded representation of the electrostatic potential of human Asf1a. The potential was calculated with the software APBS (39). The color code is red for negative values, white for near zero values, and blue for positive values. Figures were generated with PYMOL, and residues mutated in this work are labeled.

center, and surrounded by polar and charged residues such as D54 and R108 (Fig. 3).

The conformation of H3 (122–135) was investigated with an ^{15}N-labeled peptide. In the absence of Asf1a, the peptide was unfolded as revealed by the reduced dispersion of the HSQC spectrum (Fig. 4*a*) and by the Hα chemical shift index (Fig. 4*b*). Large chemical shift variations were observed upon addition of unlabeled Asf1a (1–156) (Fig. 4 *a* and *c*), and the ^{15}N-H3 (122–135) peptide adopted a helical conformation from residues 122 to 131 (Fig. 4*b*) that corresponds to the C-terminal helix α3 of H3 found in the nucleosome core particle (1). The ^{15}N{^{1}H}-nuclear Overhauser effect values of the bound peptide are consistent with a structure as rigid as the Asf1a core domain (Figs. 4*d* and 6*d*) and indicate a specific binding interaction with Asf1a. Based on titration experiments, the affinity of the peptide for Asf1a (1–156) was measured to be 100 $\pm$ 20 μM.

***In Vitro* Analysis of Asf1a Binding to the Purified Histone H3/H4 Complex.** Our NMR experiments identified a site on the surface of Asf1a that was involved in binding the 13 C-terminal residues of histone H3 and was thus likely to be involved in binding the histone H3/H4 complex (Fig. 3). We made Asf1a mutants within this site in the context of GST fusion proteins to test their ability to bind the H3/H4 complex *in vitro*. The V94R mutant was chosen because it converts a conserved exposed hydrophobic residue at the center of the H3 (122–135) binding site into a positively charged arginine residue. The D54R and R108E mutants reversed the charge of two conserved residues whose chemical shift was perturbed by the H3 peptide and that are found at either side of the hydrophobic groove containing V94. Finally, a D37R+E39R double mutant was made, affecting two acidic residues that reside within a known HirA-interacting region (16) that is outside of the putative H3-binding site to compare its effect to that of the three mutants within the H3 binding site. All these residues are highly conserved and are identical in the human and the yeast proteins (Fig. 8). We also verified that the HSQC spectra of the four Asf1a (1–156) mutants were nearly superimposable with that of the wild type (data not shown), indicating that these surface mutations do not affect their global folding and structural integrity.

We used a GST pull-down assay to study the binding of wild-type and mutant Asf1a proteins (D37R+E39R, D54R, V94R, and R108E) to histone H3 within purified chicken histone H3/H4 complex (Fig. 5*a*). This native histone complex can form $(H3\text{-}H4)_2$ tetrameric particles when assembled onto DNA (25). Wild-type Asf1a and the D37R+E39R mutant strongly bound H3 within the H3/H4 complex. In contrast, the V94R mutation nearly completely abolished H3 binding, whereas the D54R mutation, and to a lesser extent the R108E mutation, partially inhibited the binding.

The inability of the Asf1a-V94R mutant to bind histones was confirmed by two independent methods. First, we added Asf1-V94R to the ^{15}N-H3 122–135 peptide and found that it was unable to induce helical formation of the H3 peptide (Fig. 9, which is published as supporting information on the PNAS web site). This finding shows that the induced helical conformation

BIOCHEMISTRY

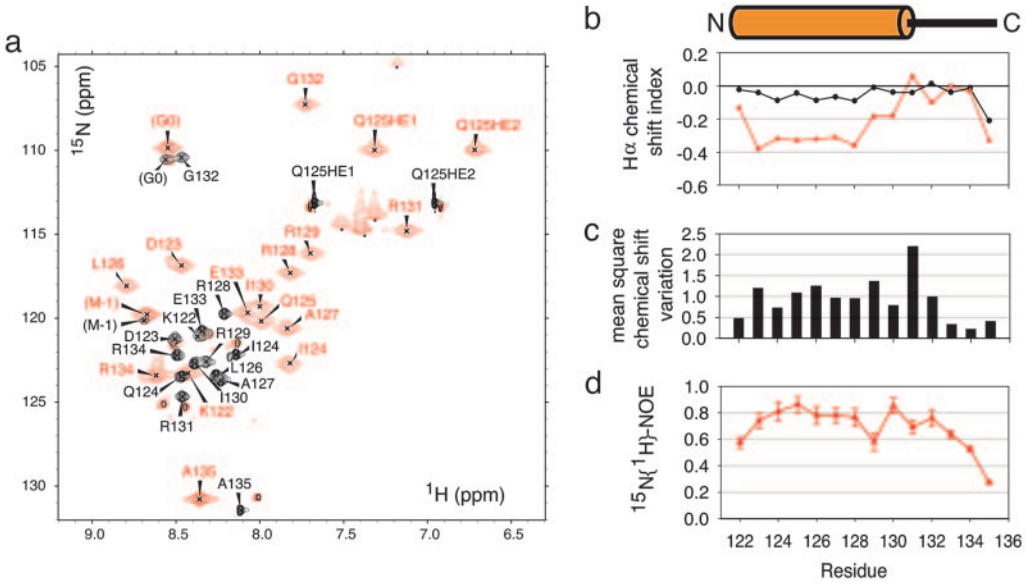


Fig. 4. H3 (122–135) folds into the native structure upon Asf1a binding. (*a*) Superposition of the ^{15}N-HSQC spectrum of ^{15}N-labeled H3 (122–135) (200 μM) with (red) and without (black) an excess of unlabeled human Asf1a (1–156) (250 μM) at 283 K. At this temperature, the chemical exchange rate was slow compared with the chemical shift variations. G_0 and M_{-1} indicate the two amino acids N-terminal to the H3 (122–135) sequence that were required for cloning the sequence in the *E. coli* expression vector as described in *Methods*. (*b*) Hα chemical shift index [$\delta_{H\alpha}^{exp}$-$\delta_{H\alpha}^{RC}$, where exp and RC refer to the experimental and random coil values, respectively, (40)] of the free (black) and bound (red) H3 (122–135). The orange cylinder shows the position of the native C-terminal α3 helix of H3 as observed in the structure of the nucleosome (1). (*c*) Mean square chemical shift variation calculated as described in *Methods* of ^{15}N-H3 (122–135) upon binding of human Asf1a (1–156) as a function of the sequence. (*d*) Heteronuclear ^{15}N{^{1}H}-nuclear Overhauser effect of the bound H3 (122–135).

of the peptide results from a highly specific interaction with the Asf1 surface identified by chemical shift mapping and centered on V94. Second, we transposed the wild type and the V94R mutant in the context of a TAP (tandem affinity purification)-tagged version of the *S. cerevisiae* Asf1 protein. Consistent with previous work (14), the histones H3 and H4 were major partners that copurified with wild-type Asf1 through tandem affinity purification. In contrast, no H3 or H4 was associated with the Asf1-V94R mutant (Fig. 5*b*). This result confirms the strong histone-binding defect of the V94R mutant.

***In Vivo* Analysis of *asf1* Mutants.** We examined the *in vivo* effects of mutations within the histone H3 or Hir binding sites of Asf1 by making V94R, D54R, R108E, D54R+R108E, and D37R+E39R mutants of the *S. cerevisiae* Asf1 protein and testing their ability to complement the phenotypes of *asf1*Δ mutants. In *S. cerevisiae*, *asf1*Δ mutants are viable, but they are temperature-sensitive for growth at 37°C (Fig. 5*d*), are highly sensitive to DNA-damaging agents that create DNA double-strand breaks (4, 13, 14), and they show defects in transcriptional silencing (9, 10, 12, 24, 28, 29). We used camptothecin and hydroxyurea as genotoxic stresses for the *asf1*Δ mutant. Camptothecin is an anticancer agent that stabilizes the covalent intermediary formed between DNA and topoisomerase I during its enzymatic relaxation of DNA (30). Collision of replication forks with these complexes leads to DNA double-strand breaks. Hydroxyurea depletes dNTP pools by inhibiting ribonucleotide reductase. Prolonged fork stalling is also thought to lead to DNA double-strand breaks (31). To examine transcriptional silencing, we used strains containing the *URA3* gene reporter inserted at three different silenced loci: the chromosome VIIL telomere (*TEL*::*URA3*) and the *HMR*a (*HMR*a::*URA3*) and *HML*α (*HML*α::*URA3*) silent mating-type cassettes. Transcriptional repression is stronger at the *HMRa* and *HML*α silent mating cassettes compared with the telomeres (32). The single *asf1*Δ mutant has only weak silencing defects, whereas *asf1*Δ *cac*Δ double mutants have strong silencing defects (9, 10, 12, 24). We thus tested transcriptional silencing in an *asf1*Δ *TEL*::*URA3* strain, as well as in *asf1*Δ *cac2*Δ double-mutant strains containing the *TEL*::*URA3*, *HMRa*::*URA3*, and *HML*α::*URA3* gene reporters. Expression of wild-type Asf1 containing a 13myc epitope fused to its carboxyl terminus from its endogenous promoter on a centromeric plasmid complemented the *asf1*Δ mutant phenotypes (Fig. 5 *d* and *e*). We then tested the function of the five Asf1 mutants in this same context. All of the mutant proteins were stably expressed in yeast at approximately the same level as the wild-type protein (Fig. 5*c*). The Asf1-V94R mutant was unable to complement any of the *asf1*Δ phenotypes. The R108E, D54R, and D54R+R108E mutants showed a graded increase in defective function. The Asf1-R108E mutant was indistinguishable from the wild-type protein in all phenotypic tests but one: it complemented the telomeric silencing defect of the *asf1*Δ *cac2*Δ double mutant slightly less than the wild-type protein (Fig. 5*e*). The Asf1-D54R mutant efficiently complemented the DNA damage sensitivity and growth defects of the *asf1*Δ mutant but appeared to function less than the wild type and Asf1-R108E mutant in all silencing strains. Finally, the D54R+R108E double mutant was clearly more defective than either single mutant in that it showed some detectable sensitivity to genotoxic stress and to growth at 37°C, as well as being totally deficient in silencing function (Fig. 5 *d* and *e*). The strong thermosensitive growth defect of the V94R mutant and the weak defect of the D45R+R108E double mutant were not due to proteolysis of the mutant proteins at 37°C (Fig. 5*c*). In conclusion, mutants defective in histone H3/H4 binding *in vitro* also showed defects in transcriptional silencing and were sensitive to DNA damaging agents and to growth at high temperatures. The severity of the phenotypes for each mutant *in vivo* was nicely correlated with the severity of their histone H3 binding defects *in vitro*.

In contrast to the H3 binding mutants, the Asf1-D37R+E39R double mutant showed silencing defects but no sensitivity to DNA damaging agents or to growth at high temperatures. These phenotypes are in agreement with those reported for a very similar Asf1-H36R+D37R double mutant. The homologous

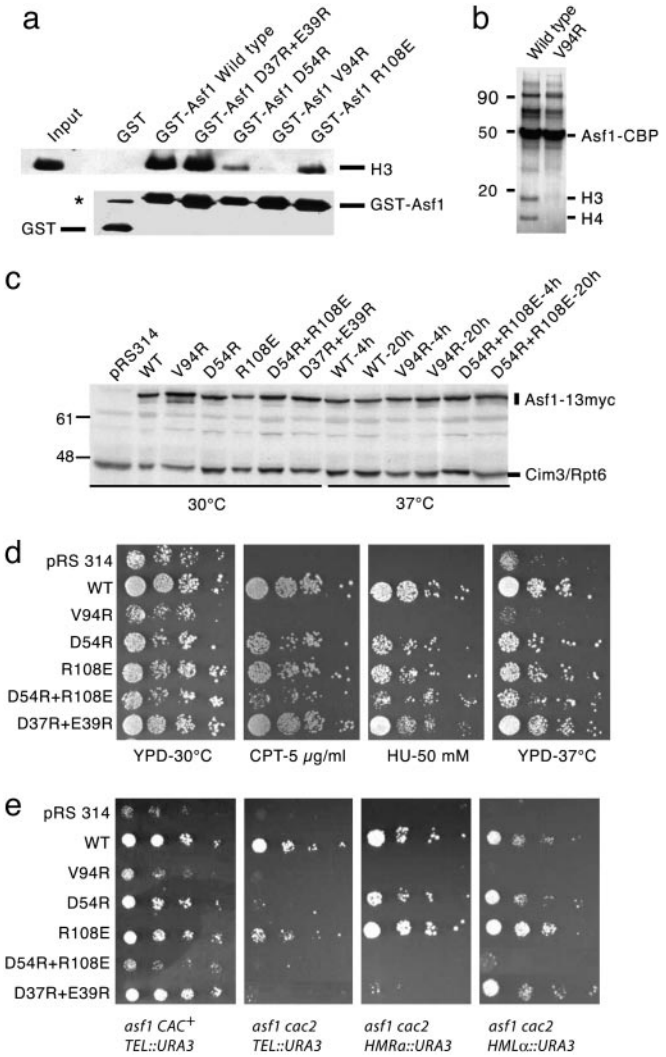


Fig. 5. *In vitro* and *in vivo* analysis of Asf1 mutants. (*a*) V94R, D54R, and R108E are involved in the binding of Asf1 to the histone H3/H4 complex. Equal amounts of purified recombinant $(His)_6$-GST, $(His)_6$-GST-Asf1a (1–156), and the indicated mutants were bound to glutathione agarose beads. *Lower* is an anti-GST immunoblot showing the equal quantities of fusion protein bound to the beads. A small quantity of the $(His)_6$-GST control protein appears as a dimerized species that migrates near the position of the $(His)_6$-GST-Asf1a proteins (indicated with a star). Binding of histone H3 within H3/H4 complexes to the GST proteins is shown in *Upper*. H3 was visualized by immunoblotting with anti-H3 antibodies. Input shows the amount of histone used for each binding reaction. (*b*) Coomassie blue-stained gel of wild-type Asf1-TAP and Asf1-V94R-TAP purified from yeast cell extracts by using the tandem affinity purification protocol. (*c*) Wild-type and mutant Asf1–13myc proteins are expressed at similar levels in yeast. The UCC6562 *asf1*Δ strain was transformed with pRS314 (*CEN-TRP1*) or pRS314 plasmids containing DNA fragments, allowing expression from the endogenous *ASF1* promoter of wild-type (WT) or mutant Asf1–13myc proteins as indicated. Protein extracts from exponentially growing cells at 30°C or from cells transferred to 37°C for 4 or 20 h as indicated were analyzed by immunoblotting with anti-myc antibodies to visualize the tagged Asf1 proteins. Asf1–13myc migrated as a closely spaced doublet of bands at ≈75 kDa. The quantity of the faster-migrating band varied between experiments, but no reproducible difference in the ratio of the two forms was seen between the wild type and any of the mutants. After developing the anti-myc blot, membranes were incubated with anti-Cim3/Rpt6 antiserum directed against a 45-kDa subunit of the 26S proteasome to verify that equal quantities of protein extract were loaded in each lane. (*d*) Sensitivity of *asf1* mutants to genotoxic stress and to growth at 37°C. UCC6562 *asf1*Δ strains transformed with the pRS314 plasmids allowing the expression of wild-type and mutant Asf1–13myc proteins described in *a* were grown to early stationary phase in synthetic complete medium without tryptophan to maintain selection for the pRS314 plasmids. The 10-fold serial dilutions of cultures were then spotted on yeast extract/peptone/dextrose plates, containing where indicated 5 μg/ml camptothecin (CPT) or 50 mM hydroxyurea (HU) and incubated at 30°C to test for sensitivity to genotoxic stress or at 37°C to test for thermosensitive growth. (*e*) The same pRS314 plasmids described above were used to test the transcriptional silencing function of the indicated wild-type and mutant Asf1–13myc proteins. These plasmids were transformed into four different silencing reporter strains: UCC6562 (*asf1 CAC+ TEL::URA3*) tests telomeric silencing in an *asf1 CAC+* background, CMY1317 (*asf1 cac2 TEL::URA3*) tests telomeric silencing in an *asf1 cac2* double-mutant background, CMY1314 (*asf1 cac2 HMRa::URA3*) tests silencing at *HMRa* in an *asf1 cac2* background, and CMY1312 (*asf1 cac2 HMLα::URA3*) tests silencing at *HMLα* in an *asf1 cac2* background. Strains were grown to early stationary phase in synthetic complete medium without tryptophan to select for the plasmid, and 10-fold dilutions of the cultures were then spotted on plates containing 5-fluoroorotic acid (5-FOA). Repression of the *URA3* reporter gene by transcriptional silencing allows growth on the FOA plates (24).

Asf1a-H36R+D37R mutant was defective in HirA binding *in vitro* (16). Interestingly, the Asf1-D37R+E39R mutant showed residual silencing function that was clearly superior to that of the Asf1-V94R and Asf1-D54R+R108E mutants in the context of the *asf1*Δ *TEL::URA3* and *asf1*Δ *cac2*Δ *HMLα*::URA3 strains. Overall, these results suggest that the Asf1-H3/H4 interaction is required for silencing at telomeres and at both silent mating cassettes.

Discussion

Characteristics of the Asf1-H3/H4 Binding Mode. We identified crucial Asf1 residues involved in the recognition of the H3/H4 complex, and we showed that the carboxyl terminus of H3 bound to Asf1 adopts a native helical conformation. Our NMR and mutational analyses suggest that the H3 C-terminal sequence is a crucial part of its interaction with the H3/H4 complex. Disrupting this interaction with the Asf1-V94R mutation prevents both interaction with the H3 C-terminal peptide and binding to the intact H3/H4 complex. Neutralization of the highly basic core histones either by using high salt concentration or adding polyacidic peptides allows assembly of core histones and DNA into nucleosome core particles *in vitro* (33, 34). It was thus reasonable to assume that electrostatic interactions would be crucial components in the interaction of physiological chaperones with core histones and, indeed, several such chaperones contain blocks of acidic sequences that may serve this role, although these acidic domains in chaperones such as Nap1, nucleoplasmin, and yeast Asf1 are not essential for their function as chaperones (16, 27, 33, 35, 36). Here, we show that the core of the Asf1a binding interface includes a conserved patch of hydrophobic residues. V94 is centrally located in this hydrophobic surface, whose chemical shifts were affected by the presence of the H3 C-terminal helix, and was also crucial for binding the intact H3/H4 complex *in vitro*. Furthermore, the yeast Asf1-V94R mutant exhibited a histone-binding defect and a strongly deleterious phenotype *in vivo*. Surrounding the hydrophobic core, charged residues such as D54 and R108 contributed to a lesser extent to the binding interface as revealed by the *in vitro* assays. A nonspecific long-range repulsive effect of the D54R mutation can be excluded because *in vitro*, the D37R+E39R double mutant affecting the same side of Asf1, binds histones H3/H4 identically to the wild type. Although the net charge of Asf1 has been maintained highly negative throughout evolution, only a few charged positions remain strictly conserved and include, in particular, the residues D54 and R108. Altogether, our results are consistent with a highly specific binding mode of Asf1 with histones H3/H4 involving hydrophobic interactions and strictly conserved charged positions.

BIOCHEMISTRY

Differential Roles for Histone H3 and Hir Protein Binding in Asf1 Function. Asf1 has been implicated in nucleosome assembly (4, 5, 8), in transcriptional silencing (9, 10, 12, 24, 28, 29, 37), and in the cellular response to DNA damage (13, 14). In yeast, it is also required for growth at high temperatures (Fig. 5) and for the transcriptional regulation of histone gene expression (9). Consistent with its multiple functions, Asf1 has been shown to physically and functionally interact with a large number of partners. The identification of these multiple interacting partners raises the question as to their role in the different pathways in which Asf1 functions. Like Asf1, many of these partners have been implicated in transcriptional activation, transcriptional silencing, and the DNA damage response. It is thus not possible to attribute specific functions to each Asf1 partner on the basis of the phenotype of individual null mutants. One approach that should allow a more precise delimitation of functions involves the identification and study of mutants that affect the interaction of Asf1 with specific partners. Recently, some mutants of Asf1a were described that decrease its binding to HirA *in vitro* (16). Analysis of one of the corresponding yeast mutants revealed that it was defective in transcriptional silencing but provided a normal level of resistance to genotoxic stress. These results thus suggested that the Asf1–Hir interaction contributes to transcriptional silencing but is not required for Asf1's role in the DNA damage response. In this study, we identified mutants of Asf1a with decreased affinity for histone H3/H4 complexes *in vitro*. Interestingly, we found an excellent correlation between the severity of the *in vitro* binding defects with that of the severity of the *in vivo* phenotypes of the corresponding yeast mutants, suggesting that binding of the histone H3/H4 complex is crucial for the role of Asf1 in the DNA damage response, in thermoresistant growth, and in transcriptional silencing, although we cannot rule out the possibility that these mutations also affect its interaction with some other important partner. Further work should provide a mechanistic description of how the histone chaperone activity of Asf1 contributes to these different functional pathways, and the identification of mutants specifically defective in binding to the remaining Asf1 partners should allow an assessment of their respective roles in these pathways.

We thank Ariel Prunell for his generous gift of purified chicken histone H3/H4 complex and for his extensive advice on histone modifications and purification, Vaughn Jackson for discussions, Dan Gottschling for the gift of silencing reporter strains, Joël Couprie, Sophie Zinn-Justin, and Bernard Gilquin for constructive discussions and precious technical support, Hervé Desvaux and Patrick Berthault for their expert NMR assistance, Carine van Heijenoort for valuable comments on the manuscript, and Geneviève Almouzni for her enthusiasm and encouragement. This work was supported by a Programme Incitatif et Coopératif of the Commissariat à l'Energie Atomique/Institut Curie on Epigenetic parameters in DNA damage response and the cell cycle, funds for the Structural Radiobiology efforts of the Common NMR Laboratory of the Commissariat à l'Energie Atomique/Saclay, and Association pour la Recherche sur le Cancer Grant 4470 (to C.M.). A.L. and B.A. are supported by a Direction Générale des Armées fellowship.

1. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997) *Nature* **389,** 251–260.
2. Loyola, A. & Almouzni, G. (2004) *Biochim. Biophys. Acta* **1677,** 3–11.
3. Verreault, A. (2000) *Genes Dev.* **14,** 1430–1438.
4. Tyler, J. K., Adams, C. R., Chen, S. R., Kobayashi, R., Kamakaka, R. T. & Kadonaga, J. T. (1999) *Nature* **402,** 555–560.
5. Mello, J. A., Sillje, H. H., Roche, D. M., Kirschner, D. B., Nigg, E. A. & Almouzni, G. (2002) *EMBO Rep.* **3,** 329–334.
6. Tagami, H., Ray-Gallet, D., Almouzni, G. & Nakatani, Y. (2004) *Cell* **116,** 51–61.
7. Adkins, M. W. & Tyler, J. K. (2004) *J. Biol. Chem.* **279,** 52069–52074.
8. Munakata, T., Adachi, N., Yokoyama, N., Kuzuhara, T. & Horikoshi, M. (2000) *Genes Cells* **5,** 221–233.
9. Sutton, A., Bucaria, J., Osley, M. A. & Sternglanz, R. (2001) *Genetics* **158,** 587–596.
10. Sharp, J. A., Fouts, E. T., Krawitz, D. C. & Kaufman, P. D. (2001) *Curr. Biol.* **11,** 463–473.
11. Tyler, J. K., Collins, K. A., Prasad-Sinha, J., Amiott, E., Bulger, M., Harte, P. J., Kobayashi, R. & Kadonaga, J. T. (2001) *Mol. Cell. Biol.* **21,** 6574–6584.
12. Krawitz, D. C., Kama, T. & Kaufman, P. D. (2002) *Mol. Cell. Biol.* **22,** 614–625.
13. Hu, F., Alcasabas, A. A. & Elledge, S. J. (2001) *Genes Dev.* **15,** 1061–1066.
14. Emili, A., Schieltz, D. M., Yates, J. R., 3rd, & Hartwell, L. H. (2001) *Mol. Cell* **7,** 13–20.
15. Sillje, H. H. & Nigg, E. A. (2001) *Curr. Biol.* **11,** 1068–1073.
16. Daganzo, S. M., Erzberger, J. P., Lam, W. M., Skordalakes, E., Zhang, R., Franco, A. A., Brill, S. J., Adams, P. D., Berger, J. M. & Kaufman, P. D. (2003) *Curr. Biol.* **13,** 2148–2158.
17. Mousson, F., Couprie, J., Thuret, J. Y., Neumann, J. M., Mann, C. & Ochsenbein, F. (2004) *J. Biomol. NMR* **29,** 413–414.
18. Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. (2003) *Bioinformatics* **19,** 315–316.
19. Nilges, M., Macias, M. J., O'Donoghue, S. I. & Oschkinat, H. (1997) *J. Mol. Biol.* **269,** 408–422.
20. Cornilescu, G., Delaglio, F. & Bax, A. (1999) *J. Biomol. NMR* **13,** 289–302.
21. Vuister, G. W. & Bax, A. (1994) *J. Biomol. NMR* **4,** 193–200.
22. Farmer, B. T., 2nd, Constantine, K. L., Goldfarb, V., Friedrichs, M. S., Wittekind, M., Yanchunas, J., Jr., Robertson, J. G. & Mueller, L. (1996) *Nat. Struct. Biol* **3,** 995–997.
23. Dubacq, C., Chevalier, A. & Mann, C. (2004) *Mol. Cell. Biol.* **24,** 2560–2572.
24. Singer, M. S., Kahana, A., Wolf, A. J., Meisinger, L. L., Peterson, S. E., Goggin, C., Mahowald, M. & Gottschling, D. E. (1998) *Genetics* **150,** 613–632.
25. Hamiche, A., Carot, V., Alilat, M., De Lucia, F., O'Donohue, M. F., Revet, B. & Prunell, A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7588–7593.
26. Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. & Seraphin, B. (2001) *Methods* **24,** 218–229.
27. Umehara, T., Chimura, T., Ichikawa, N. & Horikoshi, M. (2002) *Genes Cells* **7,** 59–73.
28. Meijsing, S. H. & Ehrenhofer-Murray, A. E. (2001) *Genes Dev.* **15,** 3169–3182.
29. Osada, S., Sutton, A., Muster, N., Brown, C. E., Yates, J. R., 3rd, Sternglanz, R. & Workman, J. L. (2001) *Genes Dev.* **15,** 3155–3168.
30. Connelly, J. C. & Leach, D. R. (2004) *Mol. Cell* **13,** 307–316.
31. Osborn, A. J., Elledge, S. J. & Zou, L. (2002) *Trends Cell Biol.* **12,** 509–516.
32. Rusche, L. N., Kirchmaier, A. L. & Rine, J. (2003) *Annu. Rev. Biochem.* **72,** 481–516.
33. Akey, C. W. & Luger, K. (2003) *Curr. Opin. Struct. Biol.* **13,** 6–14.
34. Stein, A., Whitlock, J. P., Jr., & Bina, M. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 5000–5004.
35. Arnan, C., Saperas, N., Prieto, C., Chiva, M. & Ausio, J. (2003) *J. Biol. Chem.* **278,** 31319–31324.
36. McBryant, S. J., Park, Y. J., Abernathy, S. M., Laybourn, P. J., Nyborg, J. K. & Luger, K. (2003) *J. Biol. Chem.* **278,** 44574–44583.
37. Moshkin, Y. M., Armstrong, J. A., Maeda, R. K., Tamkun, J. W., Verrijzer, P., Kennison, J. A. & Karch, F. (2002) *Genes Dev.* **16,** 2621–2626.
38. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. (2002) *Bioinformatics* **18,** S71–S77.
39. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 10037–10041.
40. Wishart, D. S., Sykes, B. D. & Richards, F. M. (1992) *Biochemistry* **31,** 1647–1651.

ARTICLE 2

Becker E, Meyer V, Madaoui H, Guerois R.
Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response.
Bioinformatics. 2006 Jun 1;22(11):1289-92.

BIOINFORMATICS DISCOVERY NOTE Vol. 22 no. 11 2006, pages 1289–1292
doi:10.1093/bioinformatics/btl075

Structural bioinformatics

Detection of a tandem BRCT in Nbs1 and Xrs2 with functional implications in the DNA damage response

Emmanuelle Becker[1,†], Vincent Meyer[2,†], Hocine Madaoui[1] and Raphaël Guerois[1,*]

[1]Service de Biophysique des Fonctions Membranaires, URA CNRS 2096, Département de Biologie Joliot-Curie and [2]Département d'Etude et d'Ingénierie des Protéines, CEA Saclay, 91191 Gif-Sur-Yvette, Cedex, France

Received on January 30, 2006; revised and accepted on February 27, 2006
Advance Access publication March 7, 2006
Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Human Nbs1 and its homolog Xrs2 in *Saccharomyces cerevisiae* are part of the conserved MRN complex (MRX in yeast) which plays a crucial role in maintaining genomic stability. NBS1 corresponds to the gene mutated in the Nijmegen breakage syndrome (NBS) known as a radiation hyper-sensitive disease. Despite the conservation and the importance of the MRN complex, the high sequence divergence between Nbs1 and Xrs2 precluded the identification of common domains downstream of the N-terminal Fork-Head Associated (FHA) domain.

Results: Using HMM–HMM profile comparisons and structure modelling, we assessed the existence of a tandem BRCT in both Nbs1 and Xrs2 after the FHA. The structure-based conservation analysis of the tandem BRCT in Nbs1 supports its function as a phosphoserine binding domain. Remarkably, the 5 bp deletion observed in 95% of NBS patients cleaves the tandem at the linker region while preserving the structural integrity of each BRCT domain in the resulting truncated gene products.

Contact: guerois@cea.fr

Supplementary information: http://www-spider.cea.fr/Groups/si6661/view.html

1 INTRODUCTION

Nbs1 in human (or Xrs2 in yeast) is an essential component of the so-called MRN complex associating Mre11, Rad50 and Nbs1 (Petrini and Stracker, 2003; van den Bosch *et al*., 2003) and plays a crucial role in DNA repair pathways (Kobayashi *et al*., 2004). The human Nbs1 protein is a 754 amino acid long protein composed of several functional domains identified from sequence analysis and biochemical experiments (Fig. 2A). At the N-terminus, a Fork-Head Associated (FHA) domain (Durocher and Jackson, 2002) followed by a single BRCA1 C-terminal (BRCT) domain (Bork *et al*., 1997; Callebaut and Mornon, 1997) can be detected from sequence to profile searches. The C-terminus of Nbs1 contains a Mre11 binding region (Desai-Mehta *et al*., 2001) and an ATM recruitment motif (Falck *et al*., 2005). In Xrs2, the *Saccharomyces cerevisiae* functional homolog of Nbs1, the FHA domain together with the Tel1 (ATM homologue) and Mre11 binding regions are conserved but the existence of a BRCT domain was never detected from sequence analysis. As a matter of fact, the sequences of Xrs2 and Nbs1 are highly divergent in the 250 amino acids following the FHA domain (10% sequence identity). Using a specific strategy, new sequences of Xrs2 homologs not present in databases such as GenBank or EMBL could be retrieved and aligned to human sequences. From the resulting multiple sequence alignment, we show that in fact two BRCT domains are present in both human Nbs1 and yeast Xrs2 right behind the FHA domain.

Tandem BRCT have been recently recognized as major mediators of phosphorylation-dependent protein–protein interactions in processes related to cell-cycle checkpoint and DNA repair functions (Glover *et al*., 2004). The ability of the tandem BRCT of Nbs1 to bind phospho-peptides was never probed before since the existence of the second BRCT was not suspected. The model-based analysis of the tandem BRCT of Nbs1 strongly suggests that it is a phosphoserine binding module. The 5 bp deletion observed in 95% of NBS patients splits up the tandem at position 218. Remarkably, this mutation preserves the structural integrity of the second BRCT at plus or minus one residue. Altogether, our findings suggest that the NBS disease could be partly linked to a disruption of the interaction properties of the tandem BRCT: cleavage of the tandem BRCT may alter the selectivity of target recognition by Nbs1 and hence affect the signaling network required for efficient DNA damage responses.

2 METHODS

An initial profile containing close homologs of Nbs1 was built from searches of the non-redundant database using PSI-BLAST (Altschul *et al*., 1997) on the MPI server (Soding *et al*., 2005). For Xrs2, the initial profile gathered three sequences retrieved from tblastn searches on the *Saccharomyces* comparative genomic database (Kellis *et al*., 2003). The profiles were enriched by aligning profiles of more divergent sequences using the profile–profile alignment method HHalign (Soding, 2005).

Iteratively, the profile–profile alignment procedure led to a global multiple sequence alignment gathering 25 sequences from human Nbs1 to *S.cerevisiae* Xrs2 (see Supplementary information). The profile consisting of 25 sequences was compared against a database of profiles built from the PDB using the HHpred server (Soding *et al*., 2005). Three structures of

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

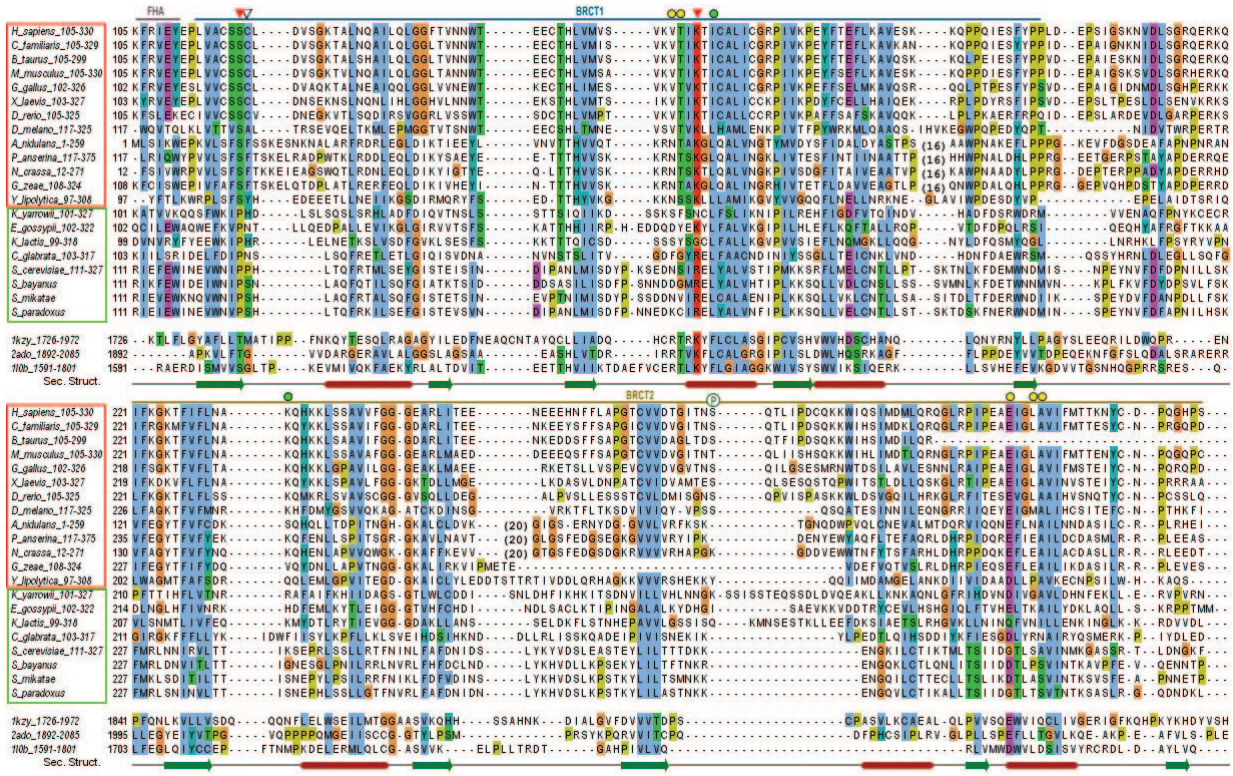


Fig. 1. Multiple sequence alignment of the tandem BRCT in homologs of Nbs1 and Xrs2 aligned with three structures of tandem BRCT (1kzy: 53BP1, 2ado: MDC1, 1l0b: BRCA1) represented with Jalview (Clamp *et al*., 2004). Helices and strands are noted by red sticks and green arrows below. Domains boundaries are shown by an horizontal line above. Red and green boxes group the sequences names with respect to their patterns at the pSer binding positions. Positions contacting the pSer residue with sidechain or backbone in tandem BRCT complex are indicated by red and white triangles, respectively. Positions contacting the pSer+3 positions are shown by green circles. Other positions found in direct contact with the phosphopeptide atoms are shown by yellow circles.

tandem BRCT were detected with significant scores and confidence levels >96% (PDB codes 1l0b, 2ado, 1kzy). Over the major length of the profile, the built alignment was consistent with the structural alignment of the templates. Yet, significant divergence could be observed at the N-terminus (first strand) and C-terminus (last $\alpha/\beta/\alpha$ motif). At the N-terminus, only the alignments with 1l0b and 2ado were compatible with the presence of an upstream FHA domain. At the C-terminus, only the alignment with the 1kzy template suggested the existence of a long insertion in the $\beta 3/\alpha 2$ loop of the second BRCT, consistent with the conservation profile in the whole Nbs1 family. A global sequence to structure alignment between the 25 sequences and the structural alignment of the three templates (1l0b, 2ado and 1kzy) was created based on these features.

Models were generated for both human Nbs1 and *S.cerevisiae* Xrs2 with Modeller 8v2 (Sali and Blundell, 1993) using the three structures 1kzy, 2ado and 1l0b as templates (max. Seq. ID: 13.2%). The quality of the models was assessed using Verify3D (Luthy *et al*., 1992), Prosa2003 (Sippl, 1993), ProQ and MaxSub (Wallner and Elofsson, 2003). The profile–profile alignment between the tandem BRCT of the Nbs1/Xrs2 family and that of the structural alignment of the three templates was iteratively refined in order to reduce the alignment errors pinpointed by the four evaluation scores.

To further assess the physical relevance of the model built for the tandem BRCT of Nbs1, a 5 ns molecular dynamic simulation was performed at 300 K in explicit solvent using GROMACS 3.2 (Van Der Spoel *et al*., 2005) (see Supplementary information for details). Conservation analyses were carried out using the Rate4site algorithm (Pupko *et al*., 2002). Possible arrangements of the FHA domain with respect to the tandem BRCT were explored using the HADDOCK program (Dominguez *et al*., 2003) by docking models of the FHA domain onto models of the tandem BRCT while constraining the distance between their C- and N-termini in respect of the Nbs1 sequence.

3 RESULTS

Models of the Nbs1 and Xrs2 tandem BRCT were built from the multiple sequence alignment in Figure 1 and assessed using standard evaluation tools. The scores of Nbs1 model are (Prosa2003: −1.92), (Verify3D: 0.395), (ProQ : 3.51) and (MaxSub : 0.348) and those of Xrs2 (Prosa2003: −1.16), (Verify3D: 0.332), (ProQ : 3.75) and (MaxSub : 0.338). The absence of residues with Verify3D scores below 0.1 together with ProQ and MaxSub scores significantly above 1.5 and 0.1, respectively, ensures the absence of major issues in the models of both tandem BRCT (Wallner and Elofsson, 2003). The physical quality of the Nbs1 model was further assessed by running a 5 ns simulation of molecular dynamics in an explicit solvent. The Cα rmsd stabilizes around 4 Å from the initial model structure (3.2 Å excluding the long loops 201–216 and 273–291) and secondary structures are overall preserved after a 5 ns of simulation as illustrated in Figure 2B (see also Supplementary information).

3.1 Functional insights from the tandem BRCT model

3.1.1 Clues for phosphoserine binding in Nbs1 So far, the tandem BRCT repeats of MDC1, PTIP, BARD1, 53BP1, RAD4, Ect2, TOPBP1, DNA ligase IV, *S.pombe* Crb2 and *S.cerevisiae* Rad9 have been shown to have phospho-serine (pSer) binding properties *in vitro* (Manke *et al*., 2003; Yu *et al*., 2003). The consensus signature for the pSer binding property was described as [S/T-G] in $\beta 1/\alpha 1$ loop

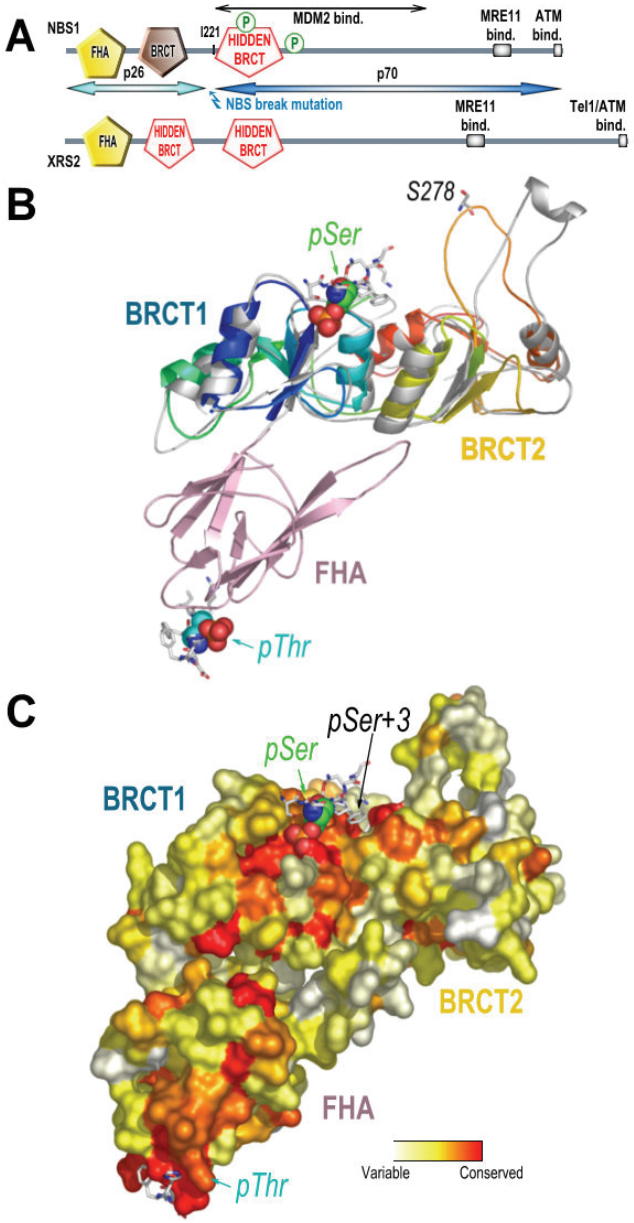


Fig. 2. (**A**) Domain organization of Nbs1 and Xrs2 with newly identified BRCT shown in red. (**B**) Ribbon representation of the model of the tandem BRCT before (rainbow colors) and after a 5 ns simulation of molecular dynamics (gray). In pink, a model of the FHA domain in a putative orientation with respect to the tandem. Phospho-peptides as found in known complexes of FHA and tandem BRCT with their ligands are shown as sticks. Phospho-Ser and -Thr residues are shown as spheres. (**C**) Surface projection of the evolutionary rates as calculated by the rate4site algorithm (Pupko *et al*., 2002). Colors from red to white report from the most conserved to the most variable positions. Drawn with pymol (DeLano, 2002).

and [S/T-X-K] in α2 helix of the first BRCT (Glover *et al*., 2004). The Gly in β1/α1 loop is quite versatile probably because only the backbone atoms at that position interacts with the pSer residue [53BP1 (1kzy in Fig. 1) has a Met and binds pSer *in vitro*]. In the tandem BRCT of Nbs1, the most conserved region localizes in the sites shown to bind the pSer residue in the structures of the tandem BRCT complexes (Stucki *et al*., 2005). The positions that directly contact the pSer with their sidechain or backbone are indicated by red and white triangles, respectively (Fig. 1). In sequences from *Homo sapiens* Nbs1 to *Yarrowia lipolytica* fungus (red box, Fig. 1), the consensus motif [S-C/F] in β1/α1 loop and [T/S-X-K] in α2 is strictly conserved supporting the function of this module as a pSer binding domain. In species ranging from *Kluyveromyces yarrowii* to *Solenodon paradoxus* (green box, Fig. 1), including Xrs2, positions binding pSer with their sidechains (red triangles) are conserved but do not match the consensus phospho-binding signature (Glover *et al*., 2004). The corresponding motif in *S.cerevisiae* Xrs2 is [P-P] in β1/α1 loop and [S-X-R] in α2 helix. Yet, several clues support that the tandem BRCT of Xrs2 might still be a pSer-binding module: (1) the BRCT domain of the ligase III shown to bind pSer peptides *in vitro* contains a Pro instead of a Ser in β1/α1 loop, as in Xrs2, (2) an Arg in α2 instead of a Lys is found in the tandem BRCT of ligase IV, also shown to be a pSer-binding module *in vitro* (Yu *et al*., 2003).

The groove at the interface between the BRCT domains is involved in the specific recognition of the residues flanking the pSer amino acid and is significantly conserved in the Nbs1 family (Fig. 2C). In structures of tandem BRCT/phosphopeptide complexes, the pSer+3 position was shown to hold much of the binding selectivity (Glover *et al*., 2004). Positions whose sidechain were shown to directly contact the position pSer+3 are indicated by green circles in Figure 1. In contrast to known structures where hydrophobic residues are often found at those positions, a Lys is quite conserved in one position of the Nbs1 multiple alignment (K233 in *H.sapiens* Nbs1).

3.1.2 Location of the phosphorylated sites in Nbs1 In response to ionizing radiation, Nbs1 is phosphorylated at Ser278 and Ser343 by the ATM kinase, and this event is required for activation of the intra S phase checkpoint (Kobayashi *et al*., 2004). From the structural model, Ser278 is located in the long β3/α2 loop of the second BRCT (Fig. 2B) and Ser343 is found 13 residues after the last residue of the tandem BRCT. Interestingly, the flexible linkers surrounding Ser278 and Ser343 are not long enough to allow for an intramolecular recognition of the pSer by the tandem BRCT.

3.1.3 Tandem BRCT and disease related mutations Of the NBS patients, 95% carry a 5 bp deletion in exon 6 of the *NBS*1 gene, which results in the expression of two truncated proteins of 26 (p26) and 70 kDa (p70) (Fig. 2A). The mutation splits the tandem precisely in the linker between the two BRCT domains. P26 moiety includes the region 1–218 spanning the FHA and the integrality of the first BRCT domain. P70 corresponds to the C-terminal half of Nbs1 and is produced by an alternative initiation of translation upstream of the 5 bp deletion. After a 18 residue extension at the N-terminus, the sequence of p70 is identical to that of the wild-type Nbs1 from I221 to the end (Williams *et al*., 2002).

I221 sharply corresponds to the beginning of the second BRCT and is the first residue fully buried in its hydrophobic core. Several structures of well-folded single C-terminal BRCT domains isolated from a tandem support that each BRCT domain can adopt its structure independently (Gaiser *et al*., 2004; Zhang *et al*., 1998). Hence, despite the severe sequence variations induced by the mutation in the linker, elements crucial for the structural integrity of the second BRCT have been preserved. It suggests that the second BRCT may not only fold independently but also hold a function important for

viability in NBS patients. Regarding the first BRCT, it has been shown that the FHA/BRCT could bind *in vitro* the histone H2AX phosphorylated by ATM (Kobayashi *et al*., 2002). Phosphorylation of H2AX at Ser129 is among the first events of the repair of double strand breaks (Lowndes and Toh, 2005). Our data suggest that the p26 fragment (Fig. 2A) may still be able to bind pSer residues in NBS cells but with a loss of binding selectivity due to the truncation of the second BRCT. This novel hypothesis would be interesting to test in the light of the results obtained on animal models of the NBS pathology (Difilippantonio *et al*., 2005; Williams *et al*., 2002).

3.1.4 Nbs1 and Mdm2 interaction Mdm2 has been extensively studied as a negative regulator of p53 tumor suppressor (Vousden and Prives, 2005). Mdm2 overexpression was recently shown to inhibit the DNA repair function of the MRN complex and this effect required the binding of Mdm2 to Nbs1 (Alt *et al*., 2005). The region 198–314 of Mdm2 was shown to associate with the MRN complex through the central region of Nbs1 221–540. This region encompasses the newly identified second BRCT domain 221–330 but not the first one. Downstream of the second BRCT, the region 330–540 is predicted to be largely unfolded (see Supplementary information). We hypothesize that the second BRCT of Nbs1 by itself may be involved in the interaction with Mdm2.

3.2 Functional implications from the FHA-tandem BRCT structural model

A striking feature of the domain organization among all Nbs1 homologs is the absence of a linker between the FHA and the tandem BRCT modules. Despite the high versatility in position and length of the insertions inside the FHA or the BRCT and between the two BRCT, not even a single amino acid was ever added at the hinge between the two modules. A structural model of the ensemble composed by the FHA and the tandem BRCT domains was built to probe the potential organization of the modules (Fig. 2B). Owing to steric hindrance, the phospho-binding sites of both domains are constrained on opposite sides of the whole assembly and could hardly be closer than 45 Å (see Supplementary information). It excludes the possibility to bind simultaneously a pThr neighboring a pSer at <15 residues. The structural constraint between the domains may originate from a specific evolutionary constraint coupling both pThr and pSer binding functions. Interestingly, the FHA and the BRCT were shown to be both required for optimal chromatin association of the MRN complex (Kobayashi *et al*., 2002; Zhao *et al*., 2002). Moreover, a mutation disrupting the FHA pThr binding site revealed that this domain is involved in a signal amplification step crucial for DNA repair after low doses of irradiation (Difilippantonio *et al*., 2005). The coupling between pThr and pSer binding functions suggested from the model might as well contribute to this amplification process.

ACKNOWLEDGEMENTS

The authors are grateful to F. Ochsenbein, M.-C. Marsolier-Kergoat and S. Zinn-Justin for their useful comments about the manuscript. This work is partly funded by the ACI IMPBIO 2004. V.M. is supported by an AFM fellowship (Association Française contre les Myopathies). H.M. is supported by a DGA fellowship. Funding to pay the Open Access publication charges was provided by the CEA Saclay.

Conflict of Interest: none declared.

REFERENCES

Alt,J.R. *et al*. (2005) Mdm2 binds to Nbs1 at sites of DNA damage and regulates double strand break repair. *J. Biol. Chem*., **280**, 18771–18781.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Bork,P. *et al*. (1997) A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *FASEB J*., **11**, 68–76.

Callebaut,I. and Mornon,J.P. (1997) From BRCA1 to RAP1: a widespread BRCT module closely associated with DNA repair. *FEBS Lett*., **400**, 25–30.

Clamp,M. *et al*. (2004) The Jalview Java alignment editor. *Bioinformatics*, **20**, 426–427.

DeLano,W.L. (2002) *The PyMOL Molecular Graphics System*. DeLano Scientific, San Carlos, CA, USA.

Desai-Mehta,A. *et al*. (2001) Distinct functional domains of nibrin mediate Mre11 binding, focus formation, and nuclear localization. *Mol. Cell. Biol*., **21**, 2184–2191.

Difilippantonio,S. *et al*. (2005) Role of Nbs1 in the activation of the Atm kinase revealed in humanized mouse models. *Nat. Cell Biol*., **7**, 675–685.

Dominguez,C. *et al*. (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc*., **125**, 1731–1737.

Durocher,D. and Jackson,S.P. (2002) The FHA domain. *FEBS Lett*., **513**, 58–66.

Falck,J. *et al*. (2005) Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. *Nature*, **434**, 605–611.

Gaiser,O.J. *et al*. (2004) Solution structure, backbone dynamics, and association behavior of the C-terminal BRCT domain from the breast cancer-associated protein BRCA1. *Biochemistry*, **43**, 15983–15995.

Glover,J.N. *et al*. (2004) Interactions between BRCT repeats and phosphoproteins: tangled up in two. *Trends Biochem. Sci*., **29**, 579–585.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kobayashi,J. *et al*. (2002) NBS1 localizes to gamma-H2AX foci through interaction with the FHA/BRCT domain. *Curr. Biol*., **12**, 1846–1851.

Kobayashi,J. *et al*. (2004) NBS1 and its functional role in the DNA damage response. *DNA Repair (Amst)*, **3**, 855–861.

Lowndes,N.F. and Toh,G.W. (2005) DNA repair: the importance of phosphorylating histone H2AX. *Curr. Biol*., **15**, R99–R102.

Luthy,R. *et al*. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.

Manke,I.A. *et al*. (2003) BRCT repeats as phosphopeptide-binding modules involved in protein targeting. *Science*, **302**, 636–639.

Petrini,J.H. and Stracker,T.H. (2003) The cellular response to DNA double-strand breaks: defining the sensors and mediators. *Trends Cell Biol*., **13**, 458–462.

Pupko,T. *et al*. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, **18** ((Suppl. 1)), S71–S77.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Sippl,M.J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins*, **17**, 355–362.

Soding,J. (2005) Protein homology detection by HMM–HMM comparison [Erratum, (2005), *Bioinformatics*, **21**, 2144.]. *Bioinformatics*, **21**, 951–960.

Soding,J. *et al*. (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res*., **33**, W244–W248.

Stucki,M. *et al*. (2005) MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell*, **123**, 1213–1226.

van den Bosch,M. *et al*. (2003) The MRN complex: coordinating and mediating the response to broken chromosomes. *EMBO Rep*., **4**, 844–849.

Van Der Spoel,D. *et al*. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem*., **26**, 1701–1718.

Vousden,K.H. and Prives,C. (2005) P53 and prognosis: new insights and further complexity. *Cell*, **120**, 7–10.

Wallner,B. and Elofsson,A. (2003) Can correct protein models be identified? *Protein Sci*., **12**, 1073–1086.

Williams,B.R. *et al*. (2002) A murine model of Nijmegen breakage syndrome. *Curr. Biol*., **12**, 648–653.

Yu,X. *et al*. (2003) The BRCT domain is a phospho-protein binding domain. *Science*, **302**, 639–642.

Zhang,X. *et al*. (1998) Structure of an XRCC1 BRCT domain: a new protein–protein interaction module. *EMBO J*., **17**, 6404–6411.

Zhao,S. *et al*. (2002) Functional analysis of FHA and BRCT domains of NBS1 in chromatin association and DNA damage responses. *Nucleic Acids Res*., **30**, 4815–4822.

ARTICLE 3

Madaoui H, Becker E, Guerois R.
Sequence search methods and scoring functions for the design of protein structures.
Methods Mol Biol. 2006;340:183-206.

9

Sequence Search Methods and Scoring Functions for the Design of Protein Structures

Hocine Madaoui[§], Emmanuelle Becker[§], and Raphael Guerois

Summary

This chapter focuses on the methods developed for the automatic or semiautomatic design of protein structures. We present several algorithms for the exploration of the sequence space and scoring of the designed models. There are now several successful designs that have been achieved using these approaches such as the stabilization of a protein fold, the stabilization of a protein–protein complex interface, and the optimization of a protein function. A rapid presentation of the methodologies is followed by a detailed analysis of two test case studies. The first one deals with the redesign of a protein hydrophobic core and the second one with the stabilization of a protein structure through mutations at the surface. The different approaches are compared and the consistency of the predictions with the experimental data are discussed. All the programs tested in these protocols are freely available through the internet and may be applied to a wide range of design issues.

Key Words: Protein design; sampling; conformational search; stability; scoring.

1. Introduction

Protein design aims at optimally selecting a protein sequence for a desired structure and function. The selection procedure depends on the ability to efficiently score a large number of protein structures. The scoring issue has been tackled by several research fields ranging from structure prediction to simulation of molecular dynamics. Yet, in contrast to structure prediction algorithms, which can deal with low-resolution models, computational design has to meet stringent precision requirements. At the other extreme, intensive computational approaches, although more precise, are often incompatible with the combinatorial problem of exploring both the sequence and the side-chain conforma-

[§]The first two authors contributed equally to this work.

From: *Methods in Molecular Biology, vol. 340: Protein Design: Methods and Applications*
Edited by: R. Guerois and M. López de la Paz © Humana Press Inc., Totowa, NJ

tional spaces. Hence, in the field of protein design, specific scoring methods have been developed to account for the balance between precision and computational efficiency. In this chapter, we present several of these approaches, first by describing the major outlines of the methods and second through a tutorial describing a step-by-step design and scoring of two proteins.

In practice, depending on the complexity of the design project, two strategies can be considered. In the first approach, close inspection of the template structure and of phylogenetic data can help designing the optimized sequence (*see* Chapters 7 and 8). Only a few positions are likely to be mutated and sequence space can be iteratively explored by hand. The structural models of the mutant have to be generated and scored independently using the appropriate energy function. This approach adapted for simple engineering purpose can be carried out with a minimum of computer resources and skills and most programs are available through the internet (**Subheadings 3.3.4.–3.3.7.**)

A more sophisticated approach consists in the use of automated design programs whose academic versions were, for the first time, recently released to the scientific community (**Subheadings 3.2.1.–3.3.3.**). They increase the scope of design projects that can be tackled but usually require the user to be at ease with computer programs running under Linux/Unix environment. Based on a given structural template, these methods explicitly model all the conformations (called "rotamers") of every position mutated during the design process. A specific energy function is used to evaluate the match of the sequence for the template. To cope with the combinatorial complexity of exploring the sequences compatible with a given template, these methods rely on efficient sampling strategies such as Monte Carlo (MC), genetic algorithm (GA), or dead-end elimination (DEE) (for a general description and comparison *see* **ref. *1***). In this chapter, two of these automated design programs are presented: RosettaDesign ***(2,3)*** and EGAD ***(4).***

Based on these methodologies, there are now many examples of successful designs that illustrate the range of applications that can be tackled. Several studies have shown that the stabilization of a protein based on a high-resolution structural scaffold can be reached using these automated methods ***(5–8)***. Far more challenging is the design of new protein folds; only one such example has been recently published ***(3)***. Probably a more accessible type of design can be the geometric idealization of an existing structure ***(9)*** or the local modification of a specific region in a protein ***(10)***. The introduction of flexibility in the backbone and between the secondary structure elements during the design process is still a delicate issue but could be addressed for specific folds family ***(11,12)***. Because folding and binding process are driven by the same rules, the global design of protein structures has also open the way to the rational design of protein-protein complex interfaces. Increasing the affinity (*see* Chapter 11)

and altering the binding specificity between partners can also be undertaken using these design programs ***(13–16)***.

In this chapter, we pay particular attention to the scoring functions that can be used to assess the quality of the design and help the selection of the optimal sequence. One of the major differences between the scoring functions is the level of heuristic terms they contain. The methods can be divided into three classes: (1) statistically based methods (SEEF) that rely on the analysis of either sequence or structure databases to transform probabilities into energies; (2) physically based methods (PEEF) that rely on the derivation of energy terms from model compounds and look for the most rigorous treatment of the basic principles of physics to calculate free energy changes; or (3) protein engineering empirically based methods (EEEF), which constitute a hybrid approach that takes advantage of thermodynamic data obtained on protein mutants and of statistical information from the databases. Here we consider each of the three approaches, analyzing their importance to the scope of protein design.

To assess the performance of the various design strategies, two test cases are analyzed in this chapter. The first one deals with the redesign of the hydrophobic core of the GB1 protein and the second with the stabilization of a protein structure (CspB) through mutations at the surface.

2. Materials

2.1. Brief Description of Programs and Downloads or URLs

2.1.1. DFire-Dmutant: EEEF Scoring Method

DMutant ***(17)*** is part of the DFire package developed in the Zhou laboratory of Biophysics and Bioinformatics at the University of Buffalo. Its goal is to predict stability changes (ΔΔG) induced by a single mutation. The mutated position and the wild-type structure need to be given. The potential used by DMutant is DFire; a distance-dependent, residue-specific, all-atom, and knowledge-based potential. The predicted free-energy change caused by a mutation is calculated assuming no structural relaxation after mutation. In the original article ***(17)***, the DMutant method was validated with a dataset of 895 large-to-small mutations (the native residue is replaced by a smaller one in the mutant). The DMutant web server is available at: http://phyyz4.med.buffalo.edu/hzhou/dmutation.html. The form is very easy to fill and the server answers very quickly.

2.1.2. PoPMuSiC: SEEF Scoring Method

PoPMuSiC ***(18)*** stands for Prediction of Protein Mutations Stability Changes. It is developed in the group of M. Rooman at the Université Libre de Bruxelles. As expressed in its name, the aim of PoPMuSiC is to predict stability changes (ΔΔG) on mutations. Although it is possible to propose several

mutations, the stability changes induced by each mutation are evaluated separately. PoPMuSiC is a database-derived potential that mixes a torsion potential and a C^{μ}–C^{μ} potential. The torsion potential describes local interactions along the sequence whereas the C^{μ}–C^{μ} potential reflects nonlocal and hydrophobic interactions. The PoPMuSiC program was validated ***(18)*** with a dataset of 344 experimentally studied mutations in seven different proteins and peptides. One example of successful experimental design was achieved based on PoPMusic predictions ***(19)***. PoPMuSiC can be used either to predict the stability changes induced by precise mutations in a structure, or to propose sequence mutations predicted to stabilize the pdb structure. PoPMuSiC web server is available at http://babylone.ulb.ac.be/popmusic. The results are returned by e-mail. The e-mail contains several files; the one that contains the results summarized is the .pdf.

2.1.3. I-Mutant and I-Mutant2.0: SEEF Scoring Method (Neural Networks)

I-Mutant and I-Mutant2.0 ***(20,21)*** are neural network-based programs developed in the group of R. Casadio at the University of Bologna. The recent development of the ProTherm database ***(22)***, collecting thermodynamic data for proteins and mutants, allowed machine-learning methods to emerge. I-Mutant is a neural network that, starting from the protein sequence and structure, classifies single amino acid substitutions into those leading to positive and those leading to negative $\Delta\Delta G$. The learning/training database and the validation one are derived from ProTherm. With a dataset of 1615 mutants (only single mutations are taken into account, *see* **ref. *20***), the accuracy reaches from 80% with I-Mutant alone to more than 90% when I-Mutant is coupled with another energy-based function such as Foldef (*see* **Subheading 2.1.5.**). I-Mutant2.0 is a support vector machine based on the same hypothesis. A web server dedicated to I-Mutant 2.0 is available at http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi (click on the "protein structure" option).

2.1.4. Scap: Conformation of Side-Chain Prediction Program

Scap ***(23)*** is part of the Jackal package developed in B. Honig's laboratory at Columbia University and the Howard Hughes Medical Institute, New York. Its first aim is to predict side-chain conformation over a rigid backbone, but it can also predict the conformation of one or more residues that have been mutated in a particular structure. Different side-chain libraries that do not depend on the backbone and different force fields are available in Scap. Scap can be found at http://trantor.bioc.columbia.edu/programs/sidechain/. The current version only supports the following platforms: SGI 6.5, Intel Linux, and Sun Solaris.

2.1.5. FoldX and Foldef Energy Function: EEEF Scoring Function

FoldX ***(24)*** is developed in L. Serrano's group at EMBL, Heidelberg. FoldX energy function, Foldef, is an empirical force field for the rapid evaluation of

the effect of mutations on the stability of proteins and nucleic acids based on its structure. The mutants structure needs to be given as an input. This is why, at this stage, FoldX is used downstream as a structure prediction method like Scap to perform design. The predictive power of Foldef has been tested on a very large set of point mutants (1088 mutants) spanning most of the structural environments found in proteins ***(24)***. Implementation of the full molecular design toolkit capabilities is in development. A web server dedicated to FoldX ***(25)*** can be found at http://foldx.embl.de/. A former version can also be found at http://fold-x.embl-heidelberg.de:1100/cgi-bin/main.cgi.

2.1.6. RosettaDesign: Sequence Search by MC and EEEF Scoring Function

RosettaDesign, developed in D. Baker's group in the University of Washington, Seattle ***(3)***, is a program aimed at finding the lowest free energy sequences for a given target protein backbone. The first application of this program is the creation of novel proteins with arbitrarily chosen three-dimensional structures. The procedure was used to design a 93-residue protein called Top7 with a novel sequence and topology (Top7 was found experimentally to be folded, and the X-ray crystal structure of Top7 is very similar to the target model). It can also be used to enhance protein stability and create alternative sequences for naturally occurring proteins. In a recent study ***(2)***, RosettaDesign was used to predict sequences with low free energy for naturally occurring protein backbones and was able to propose proteins more stable than their wild-type counterparts. A web server dedicated to RosettaDesign can be found at http://rosettadesign.med.unc.edu/. The source codes are available to academics, are located in the directory Rosetta, and can be compiled with the GNU compiler (g77 or gcc) on a computer running Linux. To compile the program, go to the Rosetta directory and type "make."

2.1.7. EGAD: Sequence Search by GA and PEEF Scoring Function

EGAD (a GA for protein Design **ref. *4***) is a protein design application developed in T. Handel's group in the University of California, Berkeley. Its main focus is to perform protein design on rigid backbone scaffolds. EGAD can perform different types of jobs such as protein designing on rigid backbones, predicting mutation effects on protein stability, or minimizing structures of proteins. The EGAD program was validated by predicting the stability of more than 1500 mutants to within 1 Kcal/mol ***(4)***.

The EGAD program is available at http://egad.berkeley.edu/software.php. The source code for EGAD is located in the directory source_code and can be compiled with the g++ compiler. To compile the program, go to the source_code directory and type "make clean all." This will create the EGAD.exe executable in the EGAD/bin directory.

2.2. Input Formats and Basic Instructions to Run the Programs

2.2.1. DFire-Dmutant

The form is very easy to fill out.

1. The native structure can be either a local pdb file or a code from the pdb (for example 1CSP for CspB or 1PGA for GB1).
2. The mutated position is indicated by its chain identifier and residue number.
3. For the mutated position, the stability change will be predicted for each of the 20 amino acids, so it is not worth indicating precisely the substitution.

DFire-Dmutant predicts stability changes induced by single mutations. As we need to test a mutant with several mutations, the best way (although not rigorous) is to test all the mutations separately and to add the predicted ΔΔG of each single mutation.

2.2.2. PopMusic

The form needs the following fields to be completed.

1. The name, e-mail, and job name: only the e-mail address is necessary for the correct execution of PoPMuSiC (the results will be sent to this address, so it must be checked carefully). The predicted stability changes induced by each single mutation are listed in the result file. To obtain a global score for the mutant, it is possible (although not exact) to add the ΔΔG of each single mutation.
2. The wild-type structure, either a local pdb file or a pdb identifier (1CSP, 1PGA).
3. To score a given mutant, choose the "specific mutations" option. A "mutation file" is then required. It lists all the mutations to score. Note that it is possible to predict the ΔΔG of a mutant with several mutations. The syntax of the "mutation file" is easy and well described. The predicted stability changes induced by each single mutation are listed in the result file. To obtain a global score for the mutant, it is possible (although not exact) to add the ΔΔG of each single mutation.

2.2.3. I-Mutant 2.0

The form needs to be filled out with this information.

1. The pdb code of the native structure (1CSP or 1PGA).
2. The chain and residue number of the mutated position.
3. For a precise substitution—for example, E3R—it is possible to specify that only mutation to arginine should be taken into account by indicating "R" in the "New Residue" field. If no "New Residue" is specified, the position will be screened with all possible amino acids.
4. Select the "Free Energy Change Value (ΔΔG)."
5. A valid e-mail address for the results to be returned.

Considering that I-Mutant 2.0 is dedicated to predict the stability change induced by single mutations, the best way to study the stability change of a multiple mutant is to add the ΔΔG of each single mutation.

2.2.4. Scap and Foldx

To predict the three-dimensional structure of the mutants, Scap needs a file that precisely lists all the mutations and all the flexible residues together with these mutations. Here are some tips to write this file.

1. To mutate a position, the syntax is "B,43,G" (residue 43 of chain B is mutated into glycine).
2. To make a position flexible, the syntax is "B,63,V" (the best rotamer for valine 63 of chain B is searched).
3. Only one mutation or one flexible residue per line.
4. The filename must end with "_scap.list."

To run Scap, the command line is:

```
> /usr/local/bin/scap -min 2 -prm 1 -ini 10 -seed 17 1CSP.pdb
                        mutant_scap.list
```

The output structure is written in a file named 1CSP_scap.pdb. To obtain other structures, change the seed number (*see* **Note 1**). To learn more about the other options, type "scap-h" or see the jackal/scap manual (*see* **Note 2**).

To score the structures, we use the FoldX web server. The server requires a login and a password to connect, either as a guest or as a registered user (registration is free). After having registered and connected to the server, do the following.

1. Upload a mutant file produced using Scap (preferentially) or using the whatif server (*see* **Note 3**) (it is possible to upload up to five structures at once).
2. Select the "Stability" option.
3. On the next window, leave unchanged the temperature, van der Waals design, or iron strength.
4. Calculate the stability (*see* **Note 4**).

2.2.5. RosettaDesign

RosettaDesign can be run in different modes that include the following.

1. Repacking side chain on a rigid backbone.
2. Redesigning on a rigid backbone.
3. Redesigning with a flexible backbone.

To run the program in the protein design mode, one needs different files in the running directory: (1) a starting pdb structure; (2) a file that specifies the location of input/output for the program, usually named "path.txt" (a default path.txt file is supplied with the Rosetta source code); (3) a file that specifies the subset of residues to redesign, named "resfile."

Thus, the desired properties for every amino acid to redesign have to be written explicitly in the resfile. All the instructions must be written in a given format (**Table 1**). For example, the following instructions specify to redesign

Table 1
Resfile Format for the RosettaDesign Program

Column	2	4–7	9–12	14–18	21–40
Description	Chain	Sequential residue number	Pdb residue number	Id	Amino acids to be used

the amino acid 3 (columns 9–12) of chain A (column 2), which is the third of the protein (columns 4–7) by using an amino acid between A, V, L, I, F, Y, or W (columns 21–40).

```
A 3 3 PIKA AVLIFYW
```

The identification information (columns 14–18) can take the following values: (1) NATAA (use native amino acid), (2) ALLAA (all amino acids allowed), (3) NATRO (use native amino acid and rotamer), (4) PIKAA (select individual amino acids), (5) POLAR (use polar amino acids), or (6) APOLA (use apolar amino acids).

The program MAKERESFILE (*see* **Note 5**) provided with Rosetta is available from the suite, and can be used to make an initial resfile with the following command line:

```
> makeresfile -p 1pga.pdb -default NATAA -resfile resfile
```

After the initial resfile is made, it can be modified by hand to specify the residues to redesign.

The command line used to redesign a protein considering a rigid backbone is:

```
> pFOLD.gnu –s 1pga.pdb -design -fixbb -resfile resfile -ndruns 1
```

The ndruns option indicates the number of design runs. The energies, structure, and sequence output are written in an output pdb file. A precise description of this file is given in the "README_output" file supplied with the program. To get a score file of the solution built with full atom scoring, the command line is:

```
> pFOLD.gnu -score -s 1pga_0001.pdb -fa_input -scorefile score
```

Considering backbone motion in protein design with RosettaDesign requires fragment files that can be obtained at: http://robetta.bakerlab.org/fragmentsubmit.jsp

Three fields need to be filled out: (1) registered username or registered e-mail address, (2) the target name, and (3) the sequence of the native structure in the fasta format (**Fig. 1**). Two fragment files are then produced.

To move the backbone and design a protein with RosettaDesign, three arguments must be used: (1) a two-letter identification code (e.g., "aa"), (2) a four-character code name for the protein (this must agree with the name of the fragment files and the name of the starting structure xxxx.pdb), and (3) the chain identification.

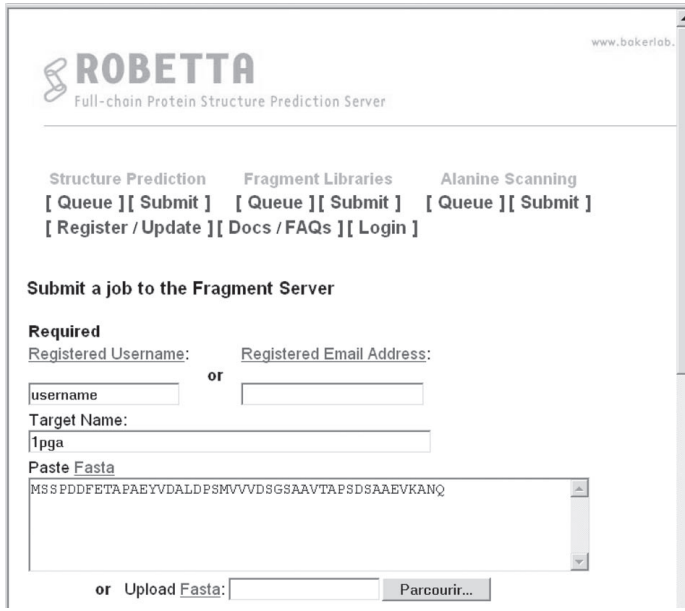


Fig. 1. Fragment files can be obtained from the Rosetta server to run RosettaDesign in flexible backbone mode.

Moreover, the starting structure must be idealized, with the instruction:

```
> pFOLD.gnu -s 1pga.pdb -idealize -fa_input
```

This command creates an output file 1pga_idl.pdb, which is used as an input to the protein design for a flexible backbone:

```
> pFOLD.gnu aa 1pga A -s 1pga_idl.pdb -design -mvbb -resfile
                                              resfile -nstruct 1
```

2.2.6. EGAD

As with RosettaDesign, the EGAD program takes user-written files as input. An EGAD input file has three sections. The first gives information about the template structure, the energy function, the desired job, and how to run it. The second section lists positions that are allowed to move. The last section, which is optional, lists positions that must be rigid during the design protocol. Here is an example of script used to redesign some positions of a starting structure spga.pdb (**Fig. 2**). The command line used to run the job is:

```
> path_to_EGAD/EGAD.exe 1pga_design
```

All the major rotamer-optimization methods (GA, MC-simulated annealing, self-consistent mean-field optimization, DEE, fast and accurate side-chain topology, and energy refinement) have been implemented in the EGAD program.

```
START
TEMPLATE_PDB ./spga.pdb
FORCEFIELD_FILE ./EGAD/examples/energy_function/forcefield
JOBTYPE mc_ga
OTHER_RESIDUES none
OUTPUT_PREFIX gb1.mc_ga
END
VARIABLE_POSITIONS
   3A  AVLIFYW
   5A  AVLIFYW
   7A  AVLIFYW
  20A  AVLIFYW
  26A  AVLIFYW
  30A  AVLIFYW
  34A  AVLIFYW
  39A  AVLIFYW
  52A  AVLIFYW
  54A  AVLIFYW
END
```

Fig. 2. Example of script spga-design used by EGAD to redesign specific positions of a starting structure template.pdb.

For our study, we have used the MC_GA method (in which solutions from MC runs are used to seed a population for genetic algorithm) (*see* **Note 6**).

2.3. pdb Used in the Example

1. Protein G (B1 immunoglobulin G-binding domain) from *Streptococcus* sp. (PDB code: 1pga.pdb).
2. Major cold shock protein (CspB) from *Bacillus subtilis* (PDB code: 1csp.pdb).

3. Methods

3.1. A Tutorial-Based Analysis of the Methods: Presentation of Two Case Studies

The β1 immunoglobulin-binding domain of streptococcal protein G (GB1) is a relevant model to study the core packing mechanisms in proteins. It has been used several times to validate design algorithms since the first studies of automatic protein design ***(8,26)***. GB1 consists of 56 residues arranged in one α-helix and two β-hairpins (**Fig. 3A**). The contiguous core of the GB1 protein is formed by 11 residues whose accessible surface area is smaller than 10% (**Table 2**). Experimentally, mutants of GB1 were found to be more stable than the wild-type GB1 ***(12)***. With three mutations in the core of the protein GB1 (**Table 3**), the midpoint of the thermal unfolding temperature can increase by up to 6°C. The stability of the second example analyzed in the chapter, the cold shock proteins (Csp), has been widely studied during the past 10 yr. They are

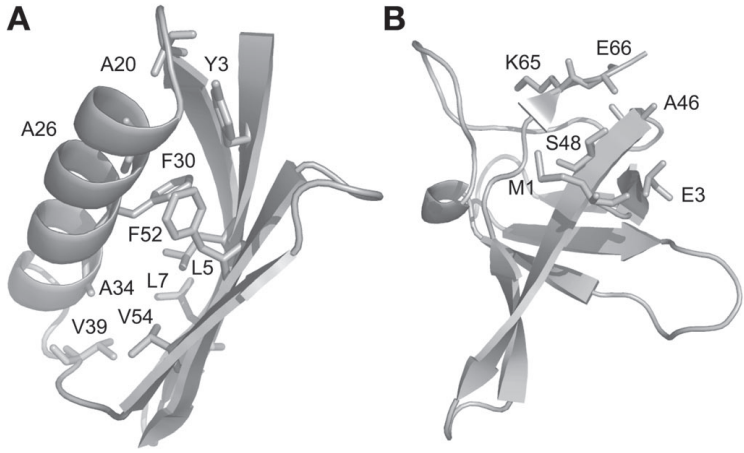


Fig. 3. Ribbon representation of the structural templates with residues mutated in the design tutorial as sticks **(A)** the GB1 protein **(B)** the CspB protein.

Table 2
Buried Residues of β1 Immunoglobulin-Binding Domain of Streptococcal Protein G (GB1)

Secondary structures	Core positions
β-sheet	3 – 5 – 7 – 20 – 43 – 52 – 54
α-helix	26 – 30 – 34
Coil	39

Their accessible surface area is <10%.

found in mesophilic, thermophilic, and hyperthermophilic bacteria, and consist of 65 to 70 residues arranged in a five-stranded antiparallel β-sheet (**Fig. 3B**). Studies proved that very few mutations at the surface of Csp proteins can lead to large differences in stability *(27)*, making them relevant examples for studying the design of protein surface. A recent article *(28)* identified stabilized variants of CspB of *Bacillus subtilis* involving the surface positions 1, 3, 46, 48, 65, and 66 using three different methods: site-specific randomization, site-directed mutations, and spontaneous mutations (error-prone polymerase chain reaction). The midpoint of the thermal unfolding transition was measured for each mutant (**Table 4**).

3.2. Sequence Search: Automatic Design Using RosettaDesign and EGAD

Our first goal was to redesign with EGAD and RosettaDesign both the core of GB1 and the surface of CspB, and to compare the sequences proposed by these programs with those known to stabilize the wild-type proteins.

Table 3
Tm is the Midpoint of Thermal Unfolding Transition (in °C)

Structures	Mutation(s)	Tm in °C
Wild-type GB1		**85.0**
Mutant GB1_1	Y3F – L7V – V39I	89.0
Mutant GB1_2	Y3F – L7I – V39I	91.0

Measures are given for the wild type β1 immunoglobulin-binding domain of streptococcal protein G and for two stabilizing mutants experimentally found (data taken from **ref. *12***).

Table 4
Tm is the Midpoint of Thermal Unfolding Transition (in °C)

Structures	Mutation(s)	Tm in °C
Wild-type CspB		**53.8**
Mutant CspB _1	N55S – Q59R	51.8
Mutant CspB _2	E43S	54.7
Mutant CspB _3	M1R – E3K – K65I	83.7
Mutant CspB _4	M1R – E3K – K65I – E66L	85.0

Measures are given for the wild type protein CspB of *Bacillus subtilis* and the four mutants studied (data taken from **ref. *28***).

In the case of GB1, positions that constitute the core, including residues 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54 were redesigned (hydrophobic amino acids A, V, L, I, F, Y, and W were considered at these positions) according to the study of Su and Mayo ***(12)***. All the other positions were kept in their native conformation.

For CspB, we tested if the protein design programs were able to propose protein variants involving the surface positions 1, 3, 46, 48, 65, and 66. Therefore, these positions were selected to be redesigned (all amino acids were considered at these positions, whereas native amino acids and rotamers were kept for other positions). To check the convergence of the programs, we did 30 different runs considering rigid or flexible backbone for every case study (*see* **Note 1**).

3.2.1. Sequence Search: Considering Rigid Backbone Design

3.2.1.1. Redesigning the Core of GB1 Using RosettaDesign or EGAD in Rigid Backbone Mode

From the different runs, only one solution is proposed by the program RosettaDesign with three mutations compared to the native structure (Y3F, L7V, F52A). These mutations were not reported in previous studies of GB1, and no experimental data can check the stabilizing effects of these mutations. We can note however that the RosettaDesign score of the produced solution (–133.42), is higher than that of the native structure (–135.27), which means

that the output solution produced considering rigid backbone is less stable than the native structure.

Using the same protocol to design the core of protein GB1 with EGAD, one sequence is proposed by the program: Y3F/V39I/V54I. As with the RosettaDesign program, the output solution (1966.96) proposed has a higher energy than the native structure (1878.56) (*see* **Note 7**).

In the study by Su and collaborators, the best protein variant found, mutant GB1_2, is more stable than the native protein. This mutant, resulting from Y3F, L7I, and V39I mutations (Tm = 91°C), has neither been proposed by the EGAD nor by the RosettaDesign program. Yet, one can note partial agreement on the Y3 and V39 positions from the EGAD results. In **Subheadings 3.3.1.** and **3.3.2.**, we analyze the scoring of the experimentally determined most stable mutants to decipher if the lack of consistency with experimental data can be explained by either a problem of conformational search or a problem in the scoring function of the programs.

3.2.1.2. Redesigning the Surface of CspB Using RosettaDesign or EGAD in Rigid Backbone Mode

With RosettaDesign, 30 different runs considering rigid backbone produce four sequences compatible with the backbone fold of the protein CspB. The EGAD program suggests one sequence (the resulting structure is less stable [2191.34] than the native structure according to the EGAD energy function [2165.92]). Unfortunately, none of the proposed sequence has been analyzed experimentally in the study of Wunderlich and collaborators ***(28)***.

3.2.2. Sequence Search: Considering Flexible Backbone Design

3.2.2.1. Redesigning the Core of GB1 Using RosettaDesign in Flexible Backbone Mode

From the different runs, four sequences compatible with the backbone fold of the protein GB1 are proposed by RosettaDesign when considering backbone motion:

1. mutant GB1_rD_1: Y3F, L7I
2. mutant GB1_rD_2: Y3F, L7V
3. mutant GB1_rD_3: L7V
4. mutant GB1_rD_4: Y3F, L7V, V39I

Several solutions are proposed emphasizing the need of launching different runs in the flexible backbone mode. For sequences 1 to 3, no experimental data are available from previous studies. However, the structure resulting from sequence 4, which was proposed by Su and Mayo in 1997, is actually more stable that the native structure of protein GB1. Indeed, the melting temperature (Tm) measured for the sequence 4 (89°C) is higher than the Tm of the native se-

quence (85°C). This example highlights the interest of running RosettaDesign with the flexible backbone mode rather than the rigid one.

3.2.2.2. Redesigning the Surface of CspB Using RosettaDesign in Flexible Backbone Mode

A total of 30 runs using flexible backbone have been done with RosettaDesign, each of them producing an output solution. From these 30 runs, 30 sequences compatible with the backbone fold of the protein CspB were proposed. However, none of the solutions was tested experimentally ***(28)***. The Shannon entropy (Hx) for every redesigned position over the 30 runs ranges from 2.31 to 3.37, which emphasizes a high variability for each position (which probably results from a problem of convergence when considering surface positions with flexible backbone).

This example highlights the difficulties of selecting a sequence when designing solvent exposed positions and stress the importance of generating multiple runs by varying the seed number (*see* **Note 1**).

3.3. Evaluating the Stabilized Protein Variants

In the previous section, the selection of sequences proposed by the RosettaDesign and EGAD were found difficult to evaluate because of the lack of experimental data validating the proposed sequences. To further assess the predictive power of the design programs, we evaluate now the scoring functions of several programs. The tests are based on both the wild-type proteins GB1 and CspB, and their respective proteins variants shown experimentally to be more stable (the mutants GB1_1, GB1_2 for GB1 and the mutants CspB_2 to CspB_4 for CspB). The CspB_1 is less stable than the wild type and has been chosen as a negative control of the prediction.

3.3.1. Scoring Design Models: RosettaDesign With Rigid Backbone

To estimate the scoring function of RosettaDesign considering rigid backbone, we have compared the score of the best mutants found for GB1 (mutants GB1_1 and GB1_2) with the wild-type protein GB1 and the score of the mutants CspB_1 to CspB_4 with the wild-type CspB.

To do so, we have set up a protocol for evaluating the mutants using the scoring function of the program, which consists of the following.

1. The wild-type protein is mutated using the WHATIF server to replace the correct atoms in the pdb (for example, we have done the Y3F, L7V, V39I mutations for the mutant GB1_1) (*see* **Note 8**).
2. Thirty different runs are launched with the mutated structure as an input.
3. A distribution of the scores from every output solution is produced in each of the run.
4. Last, every suggested mutant is ranked according to the median of its score distribution.

Table 5
Scoring of the GB1 and the CspB Optimized Protein Variants With the RosettaDesign Score Considering the Rigid Backbone Mode

Structures	RosettaDesign score	Tm (°C)
Wild-type CspB	**–111.80**	**53.8**
Mutant CspB_4	–113.86	85.0
Mutant CspB_3	–120.03	83.7
Mutant CspB_2	–124.11	54.7
Mutant CspB_1	–129.92	51.8
Mutant GB1_2	–113.60	91.0
Mutant GB1_1	–114.59	89.0
Wild-type G B1	**–135.27**	**85.0**

The lowest score refers to the most stable model. Tm is the midpoint of thermal unfolding transition (in °C).

For the evaluation of the GB1 mutants, sequence positions of the core including residues 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54 were kept identical but were allowed to move. For the evaluation of the CspB mutants, mutated residues and amino acids within 8Å of them were kept identical, but were allowed to move. In the case of GB1, the mutant GB1_2 gives a structure which is less stable than the wild-type protein when considering this protocol and the RosettaDesign rigid backbone mode. In the case of CspB, we have also ranked the mutant CspB_1 to CspB_4 and the wild-type Bs-CspB. We can conclude that the ranking of these protein variants does not agree with experimental data because all the mutants are predicted to be more stable than the wild-type protein (**Table 5**).

3.3.2. Scoring Design Models: EGAD

The same protocol as in **Subheading 3.3.1.** for evaluating the mutants (considering this time the energy function of EGAD) has been used.

We found that the mutant GB1_1 is less stable than the wild-type protein. Furthermore, the ranking of the CspB protein variants doesn't agree with experimental data because all the mutants are predicted to be more stable than the wild-type protein (**Table 6**).

3.3.3. Scoring Design Models: RosettaDesign With Flexible Backbone

The same protocol as in **Subheading 3.3.1.** for evaluating the mutants (this time running RosettaDesign in flexible backbone mode) has been used. This procedure (except from s**tep 1**) has been used also to score the wild-type proteins (*see* **Note 9**).

Launching 30 different runs in the second step is a way of exploring different conformations for a single sequence (**Fig. 4A**) (*see* **Note 1**). For example,

Table 6
Scoring of the GB1 and the CspB Optimized Protein Variants With the EGAD Energy Function

Structures	EGAD energy Kcal/mol	Tm (°C)
Wild-type CspB	**2165.92**	**53.8**
Mutant CspB_1	2107.22	51.8
Mutant CspB_2	2105.02	54.7
Mutant CspB_4	2065.78	85.0
Mutant CspB_3	2063.31	83.7
Mutant GB1_1	1886.11	89.0
Wild-type G B1	**1878.56**	**85.0**
Mutant GB1_2	1878.41	91.0

The lowest score refers to the most stable model. Tm is the midpoint of thermal unfolding transition (in °C).

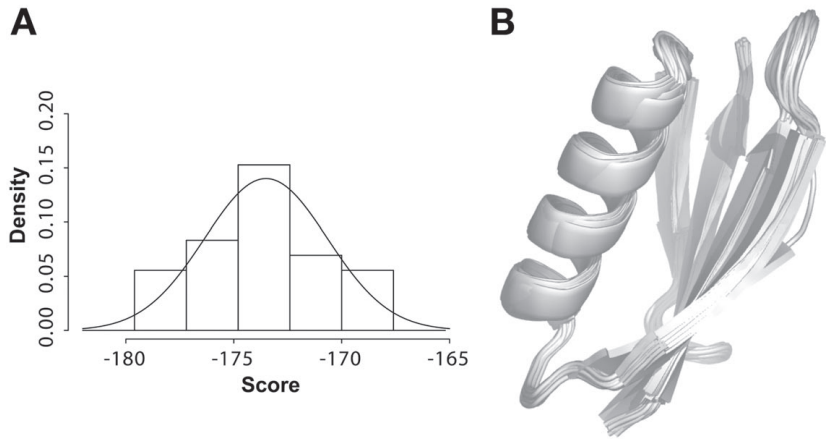


Fig. 4. Analysis of the mutant GB1_2 using RosettaDesign in flexible backbone mode. **(A)** Distribution of the RosettaDesign scores obtained with 30 different runs (varying the random seed number). **(B)** Superimposition of the ribbon representation of the 30 models, which illustrates the backbone moves explored during the design.

all the conformations generated for the mutant GB1_2 are shown on the **Fig. 4B** when considering flexible backbone with RosettaDesign (the root mean square deviation [RMSD] between all the output solutions is equal to 0.41).

This protocol has been followed for the mutants GB1_1, GB1_2, and for the wild-type protein GB1. The results are available in the **Table 7**.

According to the RosettaDesign score, the mutant GB1_1 and mutant GB1_2, which are known to be more stable than the wild-type protein, are

Table 7
Scoring of the Four Mutants of the Protein CspB and the Two Mutants of Streptococcal Protein G With the RosettaDesign Score Considering the Flexible Backbone Mode

Structures	RosettaDesign score	Tm (°C)
Mutant CspB_1	–154.12	51.8
Wild-type CspB	**–154.66**	**53.8**
Mutant CspB_2	–157.58	54.7
Mutant CspB_4	–157.74	85.0
Mutant CspB_3	–159.74	83.7
Wild-type G B1	**–172.05**	**85.0**
Mutant GB1_2	–173.67	91.0
Mutant GB1_1	–175.10	89.0

The lowest score refers to the most stable model. Tm is the midpoint of thermal unfolding transition (in °C).

predicted to stabilize the core of protein GB1. However, the mutant GB1_2 which is known to be more stable than mutant GB1_1 ***(12)*** has a lower score.

When considering flexible backbone with the RosettaDesign program, the scoring function is able to predict the stability of the protein variants with respect to the wild-type CspB. The mutant CspB_3 and the mutant CspB_4 are indeed more stable than the native protein, whereas the mutant CspB_1 is less stable (**Table 7**).

These results show a good ability of RosettaDesign to predict the stability of protein variants compared to the wild type proteins. Interestingly, good results are reported not only for the prediction of core mutations but also for mutations involving solvent exposed amino acids. These results highlight (1) the importance of considering backbone flexibility in protein design to sample a larger search space and (2) the importance of considering alternative conformations compatible with one sequence to predict the possible effect of one mutation.

3.3.4. DFire-Dmutant Scoring Function

At first, DFire-DMutant seems to have problems to deal with protein surface design of CspB (*see* **Note 10**). The mutations predicted to be the most stabilizing do not seem to be relevant.

The four mutants of CspB and the two mutants of GB1 presented previously can be scored by DFire-DMutant (**Table 8**). For the experimentally highly stabilizing mutants (*see* mutant CspB_3 and mutant CspB_4), DFire-DMutant correctly assigns a negative $\Delta\Delta G$. For the other mutants, the predictions are less reliable (as an example, *see* mutant GB1_1, experimentally found more stable than wild-type GB1, but predicted to be less stable). The quality of the

Table 8
Results Observed for the Four Mutants of the Protein CspB Previously Described and the Mutants of Streptococcal Protein G With the DFire-DMutant Program

Structures	DFire-DMutant predicted ΔΔG (kcal/mol)	Tm (°C)
Mutant CspB_2	+0.24	54.7
Wild-type CspB	**0.0**	**53.8**
Mutant CspB_1	–0.63	51.8
Mutant CspB_3	–2.38	83.7
Mutant CspB_4	–3.64	85.0
Mutant GB1_1	+0.25	89.0
Wild-type GB1	**0.0**	**85.0**
Mutant GB1_2	–0.21	91.0

The predicted ΔΔG is given in kcal/mol^{-1}. Negative ΔΔG correspond to stabilizing mutations. Tm is the midpoint of thermal unfolding transition (in °C).

Table 9
Results Observed for the Four Mutants of the Protein CspB and for the Two Mutants of Streptococcal Protein G Using PoPMuSiC

Structures	PoPMuSiC predicted ΔΔG (kcal/mol)	Tm (°C)
Mutant CspB_1	+1.03	51.8
Mutant CspB_2	+0.23	54.7
Mutant CspB_3	+0.04	83.7
Wild-type CspB	0.0	53.8
Mutant CspB_4	–0.70	85.0
Mutant GB1_1	+2.56	89.0
Mutant GB1_2	+2.40	91.0
Wild-type GB1	0.0	85.0

The predicted ΔΔG is given in kcal/mol^{-1}. Negative ΔΔG correspond to stabilizing mutations. Tm is the midpoint of thermal unfolding transition (in °C).

prediction seems to be related to the amplitude of the ΔTm between the wild-type protein and the mutants.

3.3.5. PoPMuSiC Scoring Function

The four mutants of the *B. subtilis* protein CspB and the two mutants of GB1 were submitted to PoPMuSiC (**Table 9**). For GB1, the two mutants are predicted to be significantly less stable than the wild-type CspB, although these mutants were experimentally found more stable. For CspB mutants, all mutants are predicted to be destabilizing except from one of the highly stabilizing mutant, mutant CspB_4. The quality of the predictions do not seem to correlate with the ampli-

Table 10
Results Observed for the Four Mutants of the Protein CspB and the Two Mutants of Streptococcal Protein G Using the I-Mutant2.0 Server

Structures	I-Mutant 2.0 predicted ΔΔG (kcal/mol)	Tm (°C)
Mutant CspB_1	–1.03	51.8
Wild-type CspB	**0.0**	**53.8**
Mutant CspB_3	0.19	83.7
Mutant CspB_2	0.46	54.7
Mutant CspB_4	1.49	85.0
Wild-type GB1	**0.0**	**85.0**
Mutant GB1_2	0.29	91.0
Mutant GB1_1	0.64	89.0

The predicted ΔΔG is given in kcal/mol^{-1}.By convention in I-Mutant2.0, negative ΔΔG correspond to nonstabilizing mutants, whereas positive ΔΔG correspond to stabilizing ones. Tm is the midpoint of thermal unfolding transition (in °C).

tude of the ΔTm between the wild-type protein and the mutants (for example, mutant CspB_3, the most stabilizing one, is predicted as slightly destabilizing).

3.3.6. I-Mutant2.0 Scoring Function

Table 10 summarizes I-Mutant2.0 results obtained for our two case studies (*see* **Note 11**). The sign of the ΔΔG is correct for all CspB and GB1 mutants. I-Mutant2.0 correctly identifies stabilizing and nonstabilizing mutants for both surface and core design.

3.3.7. Scap and Foldef

3.3.7.1. Scap and FoldX to Score GB1 Mutants

The two mutants of GB1 experimentally found to be more stable than the wild-type protein are Y3F-L7V-V39I and Y3F-L7I-V39I. The three mutated positions are the same and are buried (their accessible surface area is smaller than 10%).

The strategy we used to construct these mutants with Scap was the following: (1) we mutated the positions 3, 7, and 39 with amino acids F,V,I for mutant GB1_1 and F,I,I for mutant GB1_2; (2) to allow repacking in the core of the protein, we left the side chains flexible of the residues around—in a sphere of 8 Å—each mutated position; and (3) we minimized the structures with 200 steps of steepest descent using Gromos96 vacuum force field (http://www.igc.ethz.ch/gromos/) in Gromacs (http://www.gromacs.org/) to release van der Waals clashes.

Because Scap is a program that searches for a local optimum and not a global one, we recommend to do several runs of Scap with different seeds (*see* **Note 1**). We built 20 different structures for each mutant.

Table 11
Results Obtained for the Four Mutants of the Protein CspB and the Two Mutants of Streptococcal Protein G With the Strategy Combining Scap for the Prediction of Sidechain Conformations and FoldX for the Scoring Function

Structures	FoldX predicted energy—minimalist strategy	FoldX predicted energy—thorough strategy	Tm (°C)
Mutant CspB_2	18.68	19.63	54.7
Wild-type CspB	**18.11**	**17.04**	**53.8**
Mutant CspB_1	18.37	16.33	51.8
Mutant CspB_3	16.31	14.79	83.7
Mutant CspB_4	15.20	15.43	85.0
Wild-type GB1	**1.86**		**85.0**
Mutant GB1_1	1.84		89.0
Mutant GB1_2	1.14		91.0

The smaller the energy, the more stable the structure. For a mutant to stabilize the structure, its energy must be smaller than the energy of the wild type protein. Tm is the midpoint of thermal unfolding transition (in °C).

These 40 structures were scored by the FoldX server, and we selected the structure with the lowest FoldX score for each mutant (corresponding to the lowest energy). **Table 11** presents the results.

3.3.7.2. Scap and FoldX to Score CspB Mutants

The case of the CspB mutants is different from GB1 mutants because we aimed at comparing structures mutated at different positions located at the surface. To deal with the flexibility around the mutated positions (*see* **Note 12**), we propose two different strategies, a minimalist one and a thorough one: (1) mutate the positions needed, for example 55 and 59 for mutant CspB_1; (2) in the minimalist strategy, residue side chains are kept rigid around the mutated position, in the thorough strategy, all the side chains of the protein are defined as flexible; and (3) minimize the structures with 200 steps of steepest descent using Gromos96 vacuum force field (http://www.igc.ethz.ch/gromos/) in Gromacs (http://www.gromacs.org/) to release van der Waals clashes.

For each mutant, 20 structures were built and scored using the FoldX server. The lowest score for each mutant can be found in **Table 11**. The minimalist strategy and the thorough one gave quite good results. In the minimalist strategy, the highly stabilizing mutants CspB_3 and CspB_4 are predicted to be more stable than the wild-type protein with a significant energy gap (>1.5), whereas the mutants CspB_2 and CspB_1, whose Tm are close to that of the wild type are predicted to be slightly destabilizing. In the thorough strategy, the full side-

chain flexibility may account for the more disperse values. Yet, when the experimental energy gap is large, the stability of the mutants is well predicted and the hierarchy between the most stabilizing mutations is well respected.

3.4. Concluding Remarks

The results of the blind design protocols presented in this chapter show that, on average, the current algorithms offer a reliable predictive power. Design of densely packed hydrophobic regions is generally quite reliable. Stabilizing effects at the surface are more difficult to predict. Yet, when the free energy difference is large enough between the mutant and the wild type, good correlations between theory and experiments can be observed. From the global output results, the RosettaDesign software only when used with the flexible backbone option performs better. Although sequence search at the surface cannot be easily evaluated, the scoring function coupled to the flexible backbone mode gives reliable results. The support vector machine I-mutant-2.0 coupled to the Foldx program also provides interesting results in its ability to discriminate stabilizing from destabilizing mutations. It should be particularly useful for simple and fast design applications.

4. Notes

1. When running algorithms based on stochastic search methods (such as MC, GA), it is critical to run the programs several times with different random seed numbers to best explore the solutions generated by the conformational search algorithm (**Fig. 4A**). We could find dramatic variations in the interpretation of the results whether or not this repetition procedure was included in the protocols.
2. In scap "-prm 1" indicates to use Charmm22 (visit http://www.charmm.org/) with an all-atom model for the force parameters (torsion energy, van der Waals radius, or charge parameters). Other force parameters are available as "-prm 2" for Amber (visit http://amber.scripps.edu/) for all-atom model.
3. As an alternative to using Scap, the WHATIF server can be used to build simple mutations (http://swift.cmbi.kun.nl/WIWWWI/) with the menu "mutate a residue." In future version, the mutant generation facility will directly be included in the Foldx program.
4. It is possible that a pdb structure is not accepted by the FoldX server if its format is unusual. In this case, you can try to standardize the pdb file by simply opening and saving it with either Swiss PDB Viewer ***(29)*** or WHATIF ***(30)***.
5. To run these programs, make sure the most recent version of the gcc compilers is installed; otherwise, the executable file does not run properly.
6. With the version of EGAD we tested, we could not run the DEE algorithm.
7. When scoring the native structure, EGAD extracts the side-chain dihedrals and rebuilds them with ideal geometry (which causes slight deviations between the idealized and the native structure).
8. This step can be done more simply by using the PIKAA flag for every position to mutate in the resfile used by RosettaDesign.

9. The flexible backbone mode is an important feature of the RosettaDesign program that allows slight structural variations in the backbone and optimization of the solutions energy (**Fig. 4**). The studies presented in this chapter highlight the interest of activating this option. The differences between the energies of the rigid and the flexible backbone mode are huge and comparison with the wild type should be done with the same calculation conditions.
10. We noticed that the most stabilizing mutation predicted for all the positions screened is always the mutation to tryptophan. This is a frequent bias of design programs and mutations to tryptophan or other large hydrophobic residues, such as tyrosine or phenylalanine, should always be considered with caution.
11. Contrary to the methods presented in this chapter, I-Mutant and I-Mutant2.0 both predict a positive ΔΔG for stabilizing mutations. This singularity comes from the use as a training dataset of the ProTherm database that lists the free energy of unfolding instead of the free energy of folding.
12. In the case of CspB, it is not advised to restrict the flexible side chains to the residues neighboring the mutated positions. Indeed, flexible residues would be different for each mutant. Some mutants, such as CspB_3 and CspB_4, would have nearly all residues flexible, whereas mutant CspB_2 would have only a small part of its residues flexible. Because this can introduce significant noise, we propose either to keep all the neighboring residues rigid or to make all the side chains of the protein flexible.

References

1. Voigt, C. A., Gordon, D. B., and Mayo, S. L. (2000) Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **299,** 789–803.
2. Dantas, G., Kuhlman, B., Callender, D., Wong, M., and Baker, D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332,** 449–460.
3. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* **302,** 1364–1368.
4. Pokala, N. and Handel, T. M. (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J. Mol. Biol.* **347,** 203–227.
5. Filikov, A. V., Hayes, R. J., Luo, P., Stark, D. M., Chan, C., Kundu, A., and Dahiyat, B. I. (2002) Computational stabilization of human growth hormone. *Protein Sci.* **11,** 1452–1461.
6. Korkegian, A., Black, M. E., Baker, D., and Stoddard, B. L. (2005) Computational thermostabilization of an enzyme. *Science* **308,** 857–860.
7. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L., and Serrano, L. (2002) Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat. Struct. Biol.* **9,** 485–493.
8. Dahiyat, B. I. and Mayo, S. L. (1997) De novo protein design: fully automated sequence selection. *Science* **278,** 82–87.

9. Offredi, F., Dubail, F., Kischel, P., Sarinski, K., Stern, A. S., Van de Weerdt, C., et al. (2003) De novo backbone and sequence design of an idealized alpha/beta-barrel protein: evidence of stable tertiary structure. *J. Mol. Biol.* **325,** 163–174.
10. Nauli, S., Kuhlman, B., and Baker, D. (2001) Computer-based redesign of a protein folding pathway. *Nat. Struct. Biol.* **8,** 602–605.
11. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., and Kim, P. S. (1998) High-resolution protein design with backbone freedom. *Science.* **282,** 1462–1467.
12. Su, A. and Mayo, S. L. (1997) Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6,** 1701–1707.
13. Shifman, J. M. and Mayo, S. L. (2002) Modulating calmodulin binding specificity through computational protein design. *J. Mol. Biol.* **323,** 417–423.
14. Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S., et al. (2002) Computer-aided design of a PDZ domain to recognize new target sequences. *Nat. Struct. Biol.* **9,** 621–627.
15. Kortemme, T., Joachimiak, L. A., Bullock, A. N., Schuler, A. D., Stoddard, B. L., and Baker, D. (2004) Computational redesign of protein-protein interaction specificity. *Nat. Struct. Mol. Biol.* **11,** 371–379.
16. Havranek, J. J. and Harbury, P. B. (2003) Automated design of specificity in molecular recognition. *Nat. Struct. Biol.* **10,** 45–52.
17. Zhou, H. and Zhou, Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11,** 2714–2726.
18. Gilis, D. and Rooman, M. (2000) PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins. *Protein Eng.* **13,** 849–856.
19. Gilis, D., McLennan, H. R., Dehouck, Y., Cabrita, L. D., Rooman, M., and Bottomley, S. P. (2003) In vitro and in silico design of alpha1-antitrypsin mutants with different conformational stabilities. *J. Mol. Biol.* **325,** 581–589.
20. Capriotti, E., Fariselli, P., and Casadio, R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20(Suppl 1),** I63-I68.
21. Capriotti, E., Fariselli, P., and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33,** W306–W310.
22. Gromiha, M. M., Uedaira, H., An, J., Selvaraj, S., Prabakaran, P., and Sarai, A. (2002) ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. *Nucleic Acids Res.* **30,** 301–302.
23. Xiang, Z. and Honig, B. (2001) Extending the accuracy limits of prediction for side-chain conformations. *J. Mol. Biol.* **311,** 421–430.
24. Guerois, R., Nielsen, J. E., and Serrano, L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320,** 369–387.
25. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.* **33,** W382–W388.
26. Dahiyat, B. I. and Mayo, S. L. (1997) Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **94,** 10172–10177.

27. Perl, D., Mueller, U., Heinemann, U., and Schmid, F. X. (2000) Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat. Struct. Biol.* **7,** 380–383.
28. Wunderlich, M., Martin, A., and Schmid, F. X. (2005) Stabilization of the cold shock protein CspB from *Bacillus subtilis* by evolutionary optimization of Coulombic interactions. *J. Mol. Biol.* **347,** 1063–1076.
29. Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18,** 2714–2723.
30. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8,** 52–56.

ARTICLE 4

Becker E, Cotillard A, Meyer V, Madaoui H, Guérois R.
HMM-Kalign: a tool for generating sub-optimal HMM alignments.
Bioinformatics. 2007 Nov 15;23(22):3095-7.

Application note

HMM-Kalign: a tool for generating sub-optimal HMM alignments

Emmanuelle Becker*, Aurélie Cotillard, Vincent Meyer, Hocine Madaoui and Raphaël Guérois*
CEA, iBiTecS, URA 2096, SB²SM, Laboratoire de Biologie Structurale et Radiobiologie, Gif sur Yvette, F-91191 France.

ABSTRACT
Summary: Recent development of strategies using multiple sequence alignments (MSA) or profiles to detect remote homologies between proteins has led to a significant increase in the number of proteins whose structures can be generated by comparative modelling methods. However, prediction of the optimal alignment between these highly divergent homologous proteins remains a difficult issue. We present a tool based on a generalized Viterbi algorithm that generates optimal and sub-optimal alignments between one sequence and one HMM. The tool is implemented as a new function within the HMMER package called *hmmkalign*.
Availability:http://www-spider.cea.fr/Groups/hk3039/view.html
Contacts: raphael.guerois@cea.fr, emmanuelle.becker@cea.fr.

The present work aims at automatically exploring the alignment space in the neighborhood of the optimal sequence alignment (OSA) in order to find an alignment closer to the structural alignment than the OSA.

The sequence alignment space in the neighborhood of the OSA has been quite extensively explored in the context of pairwise sequence alignments. Waterman (Waterman, 1983) proposed an algorithm derived from the standard Sellers algorithm to determine all the pairwise alignments whose scores are within a range ε of the OSA's score. Later and still dealing with pairwise sequence alignments, Saqi and Sternberg (Saqi and Sternberg, 1991) proposed a heuristic known as the Iterative Elimination Method, based on the progressive perturbation of the distance matrix. Another method to generate alternative pairwise sequence alignments has been introduced by Zucker (Zuker, 1991).

With the rising of sequence-profile, sequence-HMM, and more recently profile-profile and HMM-HMM alignments, this algorithmic studies were left background. However, although progresses have been made especially for the detection of remote homology, the alignment of sequences sharing less than 25% sequence identity is still problematic in the context of comparative modelling. Based on this observation, some articles (Chivian and Baker, 2006; Jaroszewski et al., 2002; John and Sali, 2003) came back to the idea of generating alternative alignments and use heuristics such as a parametric approach (Chivian and Baker, 2006) coupled with Saqi and Sternberg's Iterative Elimination Method (Jaroszewski et al., 2002), or a genetic algorithm (John and Sali, 2003).

In this work, we explore the possibility of generating alternative alignments in the context of alignments obtained using Hidden Markov Models, such as HMMER (Eddy, 1996) or SAM (Karplus et al., 2005). Instead of heuristics, HMM-Kalign generates the exact neighborhood of the OSA.

* To whom correspondence should be addressed.

The Viterbi algorithm is classically used to align a sequence s_{obs} to a profile HMM and consists in finding the sequence of states that maximizes the emission probability of s_{obs} (Viterbi, 1967). To generate alternative alignments in the neighborhood of the OSA, one solution is to use a generalized Viterbi algorithm that precisely determines the k-best sequences of states that maximizes the emission of s_{obs}. This generalization of the Viterbi algorithm has been used in the field of speech recognition and elegant variants have been developed recently that fasten the process (Huang and Chiang, 2005). We implemented and included the generalized Viterbi algorithm in the program HMMER (Eddy, 1996).

1 GENERATING SUB-OPTIMAL ALIGNMENTS

To use the *hmmkalign* command, two files are required :

- *<MSA>*, that contains a multiple sequence alignment (derived for instance from the alignment of structural templates);
- *<sequences>*, that contains two sequences in fasta format (i) the sequence to be aligned, (ii) one sequence from the <MSA> file that may be used as a template to further build a model of the first sequence.

To build the HMM, it is possible to use the classical command :

$ *./hmmbuild <hmm file > <MSA>* (command 1)

although our results show that within highly divergent families, it is more effective to drive explicitly the HMM architecture with respect to the conservation of the secondary structures (details in supplementary data). This is possible via the command :

$ *./hmmbuild --hand <hmm file > <MSA>* (command 2)

where the *<MSA>* file contains an additional line with symbols '-' and 'x' encoding for the positions of insertions and match states, respectively. After having created the HMM, the command to generate *k* alignments is :

$ *./hmmkalign k <hmm file> <sequences>*

The OSA classically generated with HMMER corresponds to the alignment with the best score (K=1) (cf. command 1).

Exploration can be targeted to specific regions. For a sequence $s_{obs}=s_1...s_T$ in which only the region $s_i...s_j$ is to be sampled, add a hybrid sequence in the *<MSA>* file, that contains the "anchors" $s_1...s_{i-1}$ and $s_{j+1}...s_T$ and insertions '-' symbols instead of $s_i...s_j$.

2 TESTING PROCEDURE

We studied 115 alignments from 22 highly divergent protein families, *i.e.* sharing on average less than 25% identity (details in supplementary data). These alignments were extracted from the HOMSTRAD database which contains multiple structural alignments from a large set of families (Stebbings and Mizuguchi, 2004). The following procedure was applied : (1) exclude the test sequence from the multiple structural alignment; (2) build two

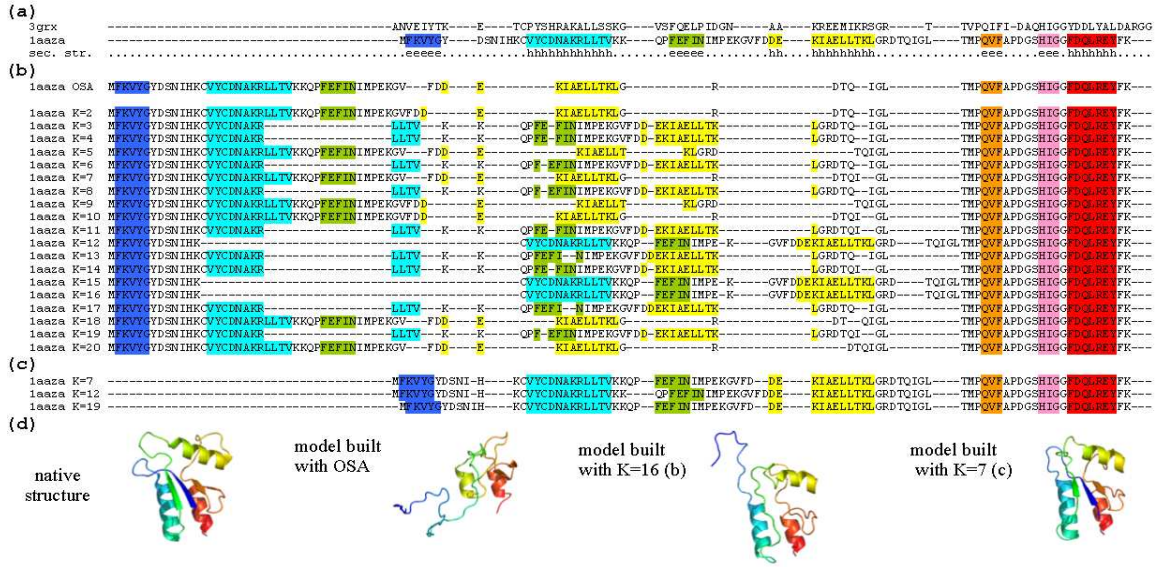


Figure 1 : Oxidized bacteriophage T4 glutaredoxin (1AAZ). The multiple alignments of the sequence with the other members of the thioredoxin family are represented through a projection into a pairwise alignment between 1aaza and 3grx. The amino acids of 1aaza corresponding to a secondary structure are highlighted in color. **(a)** The first two lines present the structural alignment and the secondary structure assigned to 1aaza (HOMSTRAD annotations). **(b)** The 20-best alignments generated when aligning the sequence over its family HMM. All alignments are different although their projection into a pairwise alignment are sometimes identical (see alignments K=12,15 and 16). Alignment K=1 corresponds to the OSA. **(c)** Alignments generated when the HMM architecture restrained by using the command 2. The 20-best alignments were computed but only 3 of them are presented (the 7th , 12th and 19th). **(d)** Native x-ray structure versus models produced by comparative modelling using the OSA, and two sub-optimal alignments, K=16 (b) and K=7 (c).

distincts HMMs with command 1 and command 2, (3) align the excluded sequence to the two HMM with hmmkalign to generate 20 alignments, (4) evaluate with respect to the structural alignment.

3 EXAMPLE WITHIN THE THIOREDOXIN FAMILY.

The thioredoxin family gathers small enzymes that are involved in redox reactions. Their sequences are about 100 amino acids long and highly divergent (17% sequence identity on average), while their 3-layer sandwich fold is conserved. Aligning the sequence of the oxidized bacteriophage T4 glutaredoxin with the other members of the family is a difficult task. As a matter of fact, the OSA (figure 1b) is far from the structural alignment (ratio of correctly aligned positions $Q_{mod} = 0.50$).

First, we studied the 20 sub-optimal alignments produced when the HMM is built with command 1 (figure 1b). The alignment can be divided in two parts: the first 63 amino acids, whose positions are extremely variable, and the last 24 amino acids that are not shifted. Interestingly, the positions that vary least along the sampled alignments correlate with the correctly aligned ones. Within the sub-optimal alignments, alignments K=12, K=15 and K=16, are substantially better than the OSA ($Q_{mod} = 0.79$).

We then studied the 20 sub-optimal alignments produced when HMM architecture is explicitly driven by secondary structure conservation (*cf.* command 2). We obtained alignments very close to the structural alignment (3 of them are presented in figure 1c), with Q_{mod} reaching 0.89.

Homology models of the oxidized bacteriophage T4 glutaredoxin were constructed with the OSA and all the sub-optimal alignments. As illustrated in figure 1d, the root mean square deviation between the native structure and the models is much smaller with models produced with the sub-optimal alignments (K=16 or K=7) than with models produced with the OSA.

4 RESULTS FOR THE 115 TEST CASES.

In 95 of the 115 test cases, there was at least one sub-optimal alignment with a better Q_{mod} than the OSA. For 26 of them, the Q_{mod} increased by more than 0.10. With respect to comparative modeling procedures, these results highlight that targeted sampling of the sequence alignment space in the neighborhood of the OSA by *hmmkalign* is efficient in generating optimized alignments and thereby better models.

REFERENCES

Chivian, D. and Baker, D. (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. *Nucleic Acids Res*, **34**, e112.

Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361-365.

Huang, L. and Chiang, D. (2005) Better k-best Parsing. *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT)*, Vancouver, BC.

Jaroszewski, L., Li, W. and Godzik, A. (2002) In search for more accurate alignments in the twilight zone. *Protein Sci*, **11**, 1702-1713.

John, B. and Sali, A. (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res*, **31**, 3982-3992.

Karplus, K., Katzman, S., Shackleford, G., Koeva, M., Draper, J., Barnes, B., Soriano, M. and Hughey, R. (2005) SAM-T04: what is new in protein-structure prediction for CASP6. *Proteins*, **61 Suppl 7**, 135-142.

Saqi, M.A. and Sternberg, M.J. (1991) A simple method to generate non-trivial alternate alignments of protein sequences. *J Mol Biol*, **219**, 727-732.

Stebbings, L.A. and Mizuguchi, K. (2004) HOMSTRAD: recent developments of the Homologous Protein Structure Alignment Database. *Nucleic Acids Res*, **32**, D203-207.

Viterbi, A.J. (1967) Error bounds for convolutional codes. *IEEE Transactions on Information Theory*, **13**, 260–269.

Waterman, M.S. (1983) Sequence alignments in the neighborhood of the optimum. *Proc Natl Acad Sci U S A*, **80**, 3123-3124.

Zuker, M. (1991) Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J Mol Biol*, **221**, 403-420.

ARTICLE 5
(soumis)

A strategy for interacting site prediction between phospho-binding modules and their partners identified from proteomic data.

Willy Aucher[1, 3,$], Emmanuelle Becker[2,4,$], Emilie Ma[1], Simona Miron[5], Arnaud Martel[6], Françoise Ochsenbein[2], Marie-Claude Marsolier-Kergoat[,1], Raphaël Guerois*[,2]*

[1] CEA, iBiTecS, SBIGeM, Laboratoire du métabolisme de l'ADN et réponses aux génotoxiques, Gif-sur-Yvette, F-91191, France.

[2] CEA, iBiTecS, SB[2]SM, Laboratoire de Biologie Structurale et Radiobiologie, Gif-sur-Yvette, F-91191, France.

[3] Present address : FRE Université/CNRS 3091, 40 Avenue du Recteur Pineau, 86022 Poitiers Cedex, France

[4] Present address : INSERM U928, Technologie Avancée pour le Génome et la Clinique (TAGC), Université de la Méditerranée, 13000 Marseille, France

[5] Institut Curie, INSERM U759, 91405 Orsay, France.

[6] CEA, iBiTecS, GIPSI, Gif-sur-Yvette, F-91191, France.

* co-corresponding authors

[$] both first authors contributed equally

Corresponding authors contact details :

Raphaël GUEROIS
Laboratoire de Biologie Structurale et Radiobiologie
iBiTecS (Institut de Biologie et de Technologie de Saclay)
Point courrier 22
CEA Saclay
91191 Gif sur Yvette cedex - FRANCE
FRANCE
tel : +33 (0)1 69 08 67 17
fax : +33 (0)1 69 08 47 12
mail : guerois@cea.fr

Marie-Claude MARSOLIER-KERGOAT
Laboratoire du métabolisme de l'ADN et réponses aux génotoxiques
iBiTecS (Institut de Biologie et de Technologie de Saclay)
Point courrier 22
CEA Saclay
91191 Gif sur Yvette cedex - FRANCE
FRANCE
tel : +33 (0)1 69 08 67 17
fax : +33 (0)1 69 08 47 12
mail : mcmk@cea.fr

28 pages of text

3 Tables

4 Figures

4 Supplementary Figures

All in all : 40 pages

Running Title : Interacting site prediction for phospho-binding modules

Summary

Small and large scale proteomic technologies are providing a wealth of potential interactions between proteins bearing phospho-recognition modules and their substrates. Resulting interaction maps reveal such a dense network of interactions that the functional dissection and understanding of these networks often require to break specific interactions while keeping the rest intact. Here, we developed a computational strategy, called STRIP, to predict the precise interaction site involved in an interaction with a phospho-recognition module. The method was validated by a two-hybrid screen carried out using the FHA1 domain of *S. cerevisiae* Rad53 as a bait, which detected two partners, Cdc7 and Cdc45, essential components of the DNA replication machinery. FHA domains are phospho-threonine binding modules and the threonines involved in both interactions could be predicted using the STRIP strategy. The threonines T484 and T189 in Cdc7 and Cdc45, respectively, were mutated and loss of binding could be monitored experimentally. The method was further tested for the analysis of 63 known Rad53 binding partners and provided several key insights regarding the threonines likely involved in these interactions. The STRIP method relies on a combination of conservation, phosphorylation likelihood and binding specificity criteria and can be accessed via a web interface at http://biodev.extra.cea.fr/strip/.

Introduction

Cell processes are tightly coordinated through signal transduction pathways that heavily depend on reversible post-translational modifications, including the phosphorylation of serine, threonine and tyrosine residues (1,2). Reflecting the multiplicity of residues being phosphorylated at a given time in a cell, a number of modules are able to mediate the specific recognition of phosphorylated partners. Typical such modules are the 14-3-3, BRCT, C2, FHA, MH2, PBD, PTB, SH2, WD-40 and WW domains (3) (see a description in dedicated databases (4,5)). These modules achieve their binding specificity primarily through the recognition of a region, usually a short sequence motifs (~6–15 residues) containing the phosphorylated residues (6).

Small and large scale proteomic technologies such as the two-hybrid technique or affinity purification are providing a wealth of potential interactions between the proteins bearing these recognition modules and their substrates. The current protein-protein interaction maps reveal a dense network of interactions with a high degree of interconnections between nodes. Deletion of large regions of a gene brings about major perturbations that complicate the functional interpretation of a specific interaction. The functional dissection of an interaction network and the understanding of its molecular logic require a more local perturbation that breaks a specific interaction while keeping the rest of the network intact. In that scope, the precise identification of the phosphorylated residue(s) responsible for an interaction often turns out to be a laborious task. Here, we propose a strategy, called STRIP (STRategy for Interacting site Prediction), to accelerate the faithful identification of these binding residues for a given phospho-binding module by coupling together several types of information : (i) the probability of a residue to be phosphorylated, (ii) the respect of the most affine motif(s) of the module around the modified

residue and (iii) the strict conservation of this motif in closely related species. The STRIP strategy can easily be used on the internet via a web server we designed for that purpose (http://biodev.extra.cea.fr/strip/).

To test this strategy we focused on one family of recognition modules, the FHA domains, and particularly addressed the case of the first FHA domain of Rad53, a *Saccharomyces cerevisiae* kinase involved in response pathways to genotoxic stresses whose catalytic domain is flanked by two FHA domains, named FHA1 and FHA2. The FHA (ForkHead Associated) domain was discovered by Hofmann and Bucher, who recognized a protein motif in a subset of forkhead-type transcription factors (7). This domain has since then been found in hundreds of proteins from eukaryotic, eubacterial and archeal species. Biochemical studies of specific FHA domains (including Rad53 domains) have demonstrated that FHA domains bind specifically phosphothreonines and have little affinity for phosphoserines or phosphotyrosines or for unphosphorylated threonines *in vitro* (8-10). Moreover, powerful *in vitro* screening strategies using combinatorial phosphopeptide libraries showed that the amino acids surrounding the phosphothreonine (pT) contribute to the FHA domains binding specificities. The highest discrimination was usually found for the amino acids in positions either (pT+3) (9,11-13) or (pT-3) (14) with a few notable exceptions (15).

The phosphopeptide binding function of the FHA domains was first demonstrated by studying Rad53 FHA1 and FHA2 (8-10) and Rad53 FHA1 domain is probably the FHA module whose biochemical and physiological characteristics have been the most thoroughly analyzed (8,16-23), which makes it an attractive target for a new predictive approach. In particular, two

groups of investigators have found that FHA1 specifically binds phosphothreonines inside pTXXD motifs *in vitro* (8,9).

Rad53 is part of the DNA checkpoints, response pathways that detect DNA lesions or replication blocks and coordinate various responses such as cell cycle arrests and transcriptional or post-translational modifications. Rad53 interacts with many different partners, and more than 30 FHA1 binding proteins have been described (18,20,24,25), although it is not always clear whether the interaction is direct or not. Abolishing FHA1 phosphopeptide binding function by mutating conserved residues like R70 and N107 leads to a slightly increased sensitivity to DNA damage from UV irradiation or MMS, but to a loss of viability on hydroxyurea (an inhibitor of ribonucleotide reductase that induces replication fork stalling), suggesting that FHA1 has a specialized function related to replicational stress (19). However, since mutating FHA1 disrupts the interactions with all its partners, the interactions involved in resistance to replicational stress remain undetermined.

In this article, we set up the STRIP strategy designed to identify the ligands bound by phosphobinding modules and we first sought to investigate FHA1 ligands, and more precisely the FHA1 ligands involved in replicational stress. We performed two-hybrid screens using Rad53 FHA1 domain as a bait to identify new partners of FHA1 or to qualify previously described interactants of Rad53 as FHA1 ligands and we isolated two essential proteins involved in DNA replication, Cdc7 and Cdc45. Using the STRIP strategy, we predicted the FHA1-bound phosphothreonines and we confirmed experimentally these predictions *in vivo* and *in vitro*. Mutating the FHA1-bound threonines of Cdc7 and Cdc45 led to no obvious phenotype, but the STRIP strategy was also able to identify as a FHA1 known ligand a threonine in Ptc2, a negative

regulator of Rad53, whose mutation led to defects in Rad53 inactivation. Finally, we applied the STRIP analysis to all Rad53 ligands.

Experimental and Computational Procedures

Plasmids

The sequence encoding Rad53 residues 1 to 164 [Rad53(1-164)] was amplified by PCR and cloned between the *EcoR*I and the *BamH*I sites of pGBT9 (Clontech), so as to create pGBT9/FHA1. The mutation of FHA1 Arg70 into alanine and the mutations of the Cdc7 and Cdc45 threonines into alanines were realized using QuickChange site-directed mutagenesis system (Stratagene). All constructs were verified by sequencing.

Two-hybrid screening

The yeast strains Y187 and Y190 were used for two-hybrid screening using the mating strategy as described in (26). We performed two-hybrid screenings using Rad53 FHA1 domain as a bait encoded by the pGBT9/FHA1 plasmid and the FRYL library of yeast genomic fragments cloned into the pACTII vector, a kind gift of Michèle Fromont-Racine and Pierre Legrain described in (Fromont-Racine, 2002, Meth Enzym). Two screenings were performed, either in the absence of genotoxic stress or in the presence of camptothecin (5 μg/ml). About 40.10^6 interactions were tested in each screening. After selection for growth on plates lacking histidine complemented with 100 mM 3-amino-triazole (3AT) and for X-Gal 5-bromo-4-chloro-3-indolyl-b-D-galactopyranoside staining, the pACTII-derived plasmids of the FRYL library were recovered and reintroduced into the testing strain in order to validate the interaction in the absence of genotoxic stress. The Cdc7 fragment was isolated in the screen performed in the presence of camptothecin but the FHA1/Cdc7 interaction was also observed in the absence of camptothecin.

Affinity measures

Phosphorylated and unphosphorylated peptides from Ptc2 (DDIpTDADTDAE), Cdc7 (DGESpTDEDDVVS) and Cdc45 (DDEApTDADEVTD) spanning the sequence of the identified FHA1 domain binding sites were obtained by chemical synthesis. The FHA1 was purified as previously described in (27). A VP-ITC Microcal was used to measure the affinities using sample concentrations (FHA1 domain (20 μM), peptides (200 μM)) at 30°C, Tris 50 mM, pH 8.

STRategy for Interacting site Prediction (STRIP)

Phosphorylation likelihood was predicted as significant if one of the two scores obtained with the NetPhos2.0 (28) and DisPhos1.3 (29) dedicated programs raised above the 0.5 threshold. The most affine binding motif recognized by the FHA1 domain of Rad53 was identified as the pTxxD motif by two independent peptide library screening studies (8,9). The conservation was analyzed from a multiple sequence alignment built with ClustalW (30) of the orthologous sequences from species closely related to *S. cerevisiae* (31,32). For every putative phosphorylated residue (threonine for the FHA) and for each residue in the position specifically recognized with respect to the phosphoresidue (+3 in the case of the Rad53 FHA1), the percentage of sequences for which the residue was strictly conserved was determined.

Results

Two-hybrid screening of Rad53 FHA1 binding proteins

We reasoned that some targets of FHA1 could bind it preferentially or exclusively in the presence of DNA damage and we performed two two-hybrid screenings using FHA1 as a bait, either in the absence of genotoxic stress or in the presence of camptothecin, an inhibitor of topoisomerase I religation reaction that induces the formation of double-strand breaks during DNA replication. Plasmid pGBT9/FHA1, encoding FHA1 (a fragment of Rad53 encompassing amino acids 1 to 164 [Rad53(1-164)]) fused to the DNA binding domain of Gal4 (Gal4BD) was used to screen a library of random genomic fragments fused to Gal4 activation domain sequence (Gal4AD). After validation, 11 proteins were found to reproducibly interact with FHA1 in the two-hybrid system (Table 1), in the absence as in the presence of camptothecin. Out of the 11 proteins, only one, Ptc2, had already been described as interacting with Rad53 FHA1 (18,33) and one, Cdc7, had been shown to be an *in vitro* phosphorylation substrate of Rad53 (34). The small overlap in the results of different screenings for protein interactants is a common observation, which in this case can be partly attributed to the fact that we used the two-hybrid technique, which mostly detects direct interactions, in contrast with affinity purification, which was used in the study of Smolka and collaborators on FHA1 partners (18).

Designing a strategy for predicting the precise interacting sites of the FHA1 domain of Rad53

The binding partners identified from two-hybrid screens or affinity-based experiments generally bear multiple residues likely to be recognized by a given phospho-recognition module. We explored whether a restricted set of residues could be isolated by screening the sequence of a

binding partner for three conditions : (i) the probability of a residue to be phosphorylated, (ii) the respect of the most affine motif around the phosphoresidue recognized by the phosphobinding module and (iii) the strict or more relaxed conservation of the most affine motif in closely related species. This strategy, called STRIP (for STRategy for Interaction site Prediction), is rationalized in more details in the following.

(i) Phosphorylation likelihood was probed using a consensus information provided by two dedicated methods, NetPhos2.0 (28) and DisPhos1.3 (29). Other efficient predictors such as GPS (35), KinasePhos (36), NetPhosK (37), PPSP (38), ScanSite (39) were not considered in this approach because they are more oriented toward the substrates of specific kinase classes such as CDK, CK2, PKA or PKC. NetPhos2.0 utilizes a neural network trained on a database of short phosphotyrosine, phosphoserine or phosphothreonine peptide fragments likely to be phosphorylated *in vivo* (28) while DisPhos1.3 algorithm is based on a logistic regression approach whose training was enhanced by including a prediction of the structural disorder and of secondary structures along the substrate sequence (29). Both methods were estimated to reach accuracies ranging between 70 and 80 % and are thus expected to provide complementary insights. (ii) The most affine motif for the FHA1 domain of Rad53 has been characterized experimentally as pTxxD. (iii) Short linear binding motifs constitute interfaces that were shown to evolve faster than globular domain-domain complexes (40). Consequently, the conservation analysis of the phosphoresidue and of its neighbouring positions was restricted to 6 fully sequenced genomes closely related to *S. cerevisiae,* namely *S. mikatae, S. paradoxus*, *S. bayanus*, *S. kluyveri*, *S. kudriavzevii* and *S. castelii* (31,32). For every putative phosphorylated residue and for each residue defined in the most affine motif (position +3 for the FHA1), the percentage of sequences for which the residue was strictly conserved was determined. If no putative phosphoresidue was detected as strictly conserved, the condition was relaxed allowing for

residues conserved in more than half the set of sequences. The rationale behind this tolerance is the possible existence of alignment flaws in long disordered regions likely to be phosphorylated and to frequent truncations in the sequences of the 6 *Saccharomyces* genomes (see discussion).

Analysis of the threonine targeted by FHA1 in Cdc7

Two two-hybrid hits, Cdc7 and Cdc45, appeared as plausible ligands for mediating FHA1 part in resistance to replication stress and were further analyzed. Cdc7 is the catalytic subunit of a kinase required for origin firing and replication fork progression. Cdc45 is a DNA replication initiation factor recruited to pre-replicative complexes at replication origins and also required for replication elongation. It has to be noted that both Cdc7 and Cdc45 are essential proteins, which precludes the analysis of deletant strains and makes compulsory the design of point mutations. We applied to the analysis of Cdc7 and Cdc45 the STRIP strategy described above.

The fragment of Cdc7 identified from the two-hybrid experiment spans the segment 294-493 and bears 9 out of the 24 threonines present in Cdc7, with 4 TxxD motifs in the fragment (Figure 1A). Only one threonine, T484, fulfilled all three criteria with high phosphorylation probability (88% and 55% according to the NetPhos and Disphos predictors, respectively), respect of the most affine TxxD motif and strict conservation of the Thr and Asp residues among closely related species. In the Cdc7(294-493) fragment, another threonine, T298, respects the TxxD motif and is strictly conserved, but has a low phosphorylation likelihood (about 17 %). We tested our *in silico* prediction by mutating T298 and T484 into alanine. As shown in Figure 2A, both the wild-type Cdc7(294-493) fragment and the Cdc7(294-493)T298A mutant interacted in the two-hybrid assay with the wild-type FHA1 domain but not with a mutant FHA1 affected in its phosphopeptide binding function (FHA1R70A). In contrast, mutating T484 into alanine

abolished Cdc7(294-493) interaction with FHA1. We verified that the wild-type and the mutant Gal4AD-Cdc7(294-493) fusions were expressed to the same levels (data not shown). Our results thus indicate that Cdc7 threonine T484 should be the target of FHA1 and validate our prediction concerning the identity of FHA1 ligand.

Interestingly, even considering the full-length Cdc7 rather than the Cdc7(294-493) fragment, we would have reached a similar conclusion since T484 was the highest scoring residue of all Cdc7 threonines.

Analysis of the threonine targeted by FHA1 in Cdc45

We identified the Cdc45(154-270) fragment as one of FHA1 interacting substrates in our screen. Cdc45 bears 33 threonines, 6 of which are located between amino acids 154 and 270 (Figure 1B). None of the 6 threonines fulfilled the three stringent criteria altogether and stringency on the conservation was relaxed in a second step as stated in the description of the STRIP methodology. Then, only one threonine out of the 6, T189, was found to fulfill the binding site criteria with a moderate conservation in 3 out of the 5 available sequences. Analysis of the multiple sequence alignment around T189 showed that it is located in a poly-acid stretch, likely disordered and difficult to align. A rapid inspection of the sequences lacking the TxxD conservation revealed that this motif could easily be identified in the neighbourhood and realigned without disrupting the alignment consistency (Supp Figure 1). To validate our prediction three point mutants of Cdc45 were designed. Selected threonines corresponded either to T189 that fulfilled all three criteria or to T245 and T195 that fulfilled only two of them. As shown in Figure 2B, the wild-type Cdc45(154-270) fragment and the Cdc45(154-270)T245A and Cdc45(154-270)T195A mutants interacted similarly in the two-hybrid assay with the wild-type FHA1 domain (and not with the mutant FHA1R70A domain). Conversely, the interaction

between FHA1 and Cdc45(154-270)T189A was weaker and we verified that this was not due to a defective expression of the Gal4AD-Cdc45(154-270)T189A fusion (data not shown). Although other parts of Cdc45(154-270) seem to participate in its interaction with FHA1, our results indicate that Cdc45 threonine T189 represents a ligand of FHA1 and validate our *in silico* predictions.

We would again have reached a similar conclusion considering the full-length Cdc45 rather than the Cdc45(154-270) fragment since T189 was one of the two highest scoring residues of all Cdc45 threonines. For the other residue, T147, close inspection of the alignment showed that the non-conservation of the TxxD motif in *S. kluyveri* could not be explained by alignment flaws or sequence truncations as for T189 (Supp Figure 1).

Rad53 FHA1 binds *in vitro* to phosphopeptides encompassing Cdc7 T484 and Cdc45 T189

In order to confirm the interactions between Rad53 FHA1 domain and Cdc7 and Cdc45, the direct binding of FHA1 to the phosphothreonine peptides ^{480}DGESpTDEDDVVS [pT(Cdc7)] and ^{185}DDEApTDADEVTD [pT(Cdc45)] derived from Cdc7 and Cdc45 sequences, respectively, was probed using Isothermal Titration Calorimetry (ITC). The dissociation constant, K_D, between Rad53 FHA1 and the pT(Cdc7) and the pT(Cdc45) reached 1.7 µM and 400 nM, respectively (Table 2 and Figure 3). pT(Cdc45) is the peptide with the highest affinity described so far for an FHA1 substrate (the peptide corresponding to the near-optimal binding motif determined for Rad53p FHA1 by peptide library screening bound with a K_D of 780 nM). These data clearly indicate that Rad53 FHA1 interacts directly with peptides encompassing Cdc7 T484 and Cdc45 T189 *in vitro* and support our hypothesis that similar, direct interactions occur *in vivo* between FHA1 and Cdc7 and Cdc45.

Mutating Cdc7 T484 and Cdc45 T189 induces no obvious replication phenotype

Having identified Cdc7 T484 and Cdc45 T189 as probable ligands of FHA1, we sought to assess the part played by the FHA1/Cdc7 and FHA1/Cdc45 interactions by abrogating specifically these interactions via the mutation of Cdc7 T484 and Cdc45 T189 into alanine. We constructed yeast strains deleted for either *CDC7* or *CDC45* at their chromosomal loci and complemented with plasmids harboring either a wild-type or a mutated copy of the corresponding gene (*cdc7T484A* and *cdc45T189A*, respectively). The strains were tested for their growth on solid medium in the presence or in the absence of various genotoxic stresses including UV-irradiation, camptothecin, hydroxyurea and 4-nitroquinoline 1-oxide (4-NQO, a reagent that produces bulky base damage of the type that is mainly repaired by the nucleotide excision repair system). No reproducible difference of viability or growth rate could be observed between the mutated *cdc7T484A* and *cdc45T189A* cells and the controls (data not shown). The double mutant *cdc7T484A cdc45T189A* also behaved as wild-type cells (data not shown). These results can be explained by the redundancy of interactions linking two proteins or even two complexes via different protein-protein interactions. Regarding Cdc45, the interaction between Rad53 FHA1 and the Cdc45T189A could be indirectly maintained in the pre-replication complex through other Rad53 partners such as Mrc1 (18) and Cdc46 (Mcm5) (41), and in the case of Cdc7, an interaction between FHA1 and Cdc7T484A could be indirectly maintained via other proteins such as Dbf4, the regulatory subunit of the Cdc7/Dbf4 kinase complex, also described as a Rad53 FHA1-mediated binding partner (24).

Application of STRIP strategy to a complex and large interactome

Cdc7 and Cdc45 represent two examples for which our strategy correctly determined the phosphothreonines targeted by FHA1 (as monitored by the complete or partial loss of interaction

in the two-hybrid test). We had previously demonstrated experimentally that FHA1 binds the threonine T376 of the PP2C phosphatase Ptc2 which plays a part in Rad53 inactivation after double-strand breaks (27). We tested whether we could have predicted this site with the STRIP strategy and found that indeed T376 is the only Ptc2 threonine fulfilling the three criteria of our test. In this case, we had demonstrated that the T376A mutation not only abolished the interaction between Ptc2 and Rad53 FHA1 but also induced a clear phenotype in terms of adaptation defects (27).

To further challenge the interest of the STRIP strategy in facilitating the dissection of large interactomes, we analyzed the whole set of physical interactions involving Rad53, either from the present work or from the literature. A number of experimental works were devoted to unravel Rad53 binding partners using either affinity-based or two hybrid-based methods (18,25,33,42). Several large scale yeast interactome analyses also provided a wealth of data connected to Rad53 (41,43-45). The graph in Figure 4 reports the 63 Rad53 binding partners extracted from this work and the Biogrid database (46) using the Osprey visualisation tool (47). Blue and red linkages report for the affinity-based and the two-hybrid results, respectively. The 11 proteins identified from our FHA1 two-hybrid screen are labelled by an obelisk (‡) in Figure 4 and in Table 3. Among the affinity-based results, 30 binding proteins (labelled by an asterisk in Figure 4 and in Table 3) were identified in a proteomic survey that focused on the isolated FHA1 domain partners (18). These interactions were lost upon point mutation of the phosphobinding site in the FHA1 domain, confirming that a phosphothreonine is mediating the interaction. For the remaining proteins, it is not known whether the FHA1, the FHA2 or another region of Rad53 is involved in the interaction. In the following, we limited our survey to the existence of putative threonines that may be recognized by FHA1 and explored how the STRIP strategy may restrict the numbers of putative binding sites.

All in all, there are 2922 threonines in the 63 binding proteins. Applying our protocol led to a unique candidate threonine for 25 out of 63 partners (Table 3). For 11 additional cases, a limited set of two to three threonines could be proposed. For the remaining 27 proteins, no threonines fulfilled the set of constraints applied on the sequence of the binding partners. These partners could bind the FHA2 or another region of Rad53, or could bind Rad53 indirectly via the intermediate of other Rad53 bridging partners. Indirect interactions concern complexes bearing many cross interactions such as the septin complex (containing Cdc3/10/11/12, Shs1, Bud4), the G1/S transition complex (made of Swi4, Swi6, Mbp1, Whi5) or the histone complex (Hht1, Hhf1, Hta2, Hmo1), for which only 5 out of 14 partners contain a candidate threonine. We wondered whether the STRIP strategy could help identifying in these stable complexes the most likely direct partner.

For the septin complex (Cdc3/10/11/12, Shs1, Bud4 network) which was found to interact with the isolated FHA1 domain (18), three proteins (Shs1, Bud4 and Cdc11, with dashed underline in Figure 4) were detected as harbouring 5 putative FHA1 binding threonines using the relaxed conservation criterion (none were found with the stringent one (see Methods)). As for Cdc45 a rapid inspection of the alignment around these 5 threonines (Supp Figure 2) revealed that only for Shs1, the limited conservation of the most affine binding site among *Saccharomyces* species was due to a sequence truncation. For the four other threonines, it corresponded unambiguously to one or several mutations, decreasing the potential functional importance of these sites. Hence, out of the 223 threonines in the proteins of the septin network, the STRIP analysis was stringent enough to restrict the binding substrate hypotheses to one partner (Shs1) and to only one residue, T539. Interestingly, Shs1 is phosphorylated by Rad53 *in vitro* (18) and appears as the only member of the septin complex to undergo a Rad53-dependent phosphorylation after treatment with the genotoxic agent methyl methane sulfonate (48).

Moreover, the morphology defect caused by overexpression of the FHA1 domain is suppressed by the deletion of *SHS1* (18), strengthening the hypothesis of a direct interaction between the two proteins.

For the G1/S transition complex (the Swi4/Swi6/Mbp1/Whi5 network) only two threonines, T64 in Swi4 and T111 in Swi6, were detected as potential binding sites, again with the less stringent conservation criterion. We can notice that FHA1 optimal binding site is more conserved around Swe6 T111 than around Swe4 T64 (Supp Figure 3), although the distinction is not as clear as for the septin complex. Interestingly, Swi6 was shown to be directly phosphorylated at residue S547 by Rad53 impacting on the delay of the G1/S transition after DNA damage (49). The STRIP analysis suggests that a direct interaction between the two proteins may be mediated by the FHA1 domain bound to a Swi6 phosphothreonine at position T111.

Other interesting features of Figure 4 coupled to the STRIP predictions are that the threonines T346 and T247 which are likely to be bound by FHA1 in Ifh1 and Yta7, respectively, are also predicted to be phosphorylated by the CK2 kinase according to the NetPhosK server (37). Interestingly, both proteins were found to associate with the Ckb2 CK2 regulatory subunit (44,50).

Discussion

With the development of large scale phospho-proteomic experiments, a number of methods have been developed to predict and analyze the phosphorylation patterns in protein substrates. The STRIP strategy specifically addresses an issue which was not considered before by combining knowledge from the bait (phospho-binding modules bearing a specific consensus binding motif) and the preys (phosphorylation likelihood and conservation of the consensus motif). It offers stringent and efficient clues to dissect the interaction networks mediated by phosphobinding protein modules. Figure 1 and the number of TxxD motifs in Table 3 illustrate that prediction of the interaction sites within FHA1 binding partners solely based on the search for the most affine motif would lead to many more threonine candidates. In the case of the FHA1 domain, the binding specificity is known from targeted experiments but recent computational approaches, such as D-MIST, suggest that these specificities may shortly be inferred from prediction (51).

To estimate the phosphorylation likelihood, the STRIP strategy relies on a meta-prediction approach combining two different algorithms NetPhos and DisPhos. However, the precision of these approaches may still be questionable and somehow the conservation of the phosphoresidue strengthens or decreases the reliability of the phosphorylation prediction. However, a characteristic feature of the phosphorylated regions is that they are often located in disordered regions which may reveal tricky to align properly. Moreover, the simple linear organization of the binding motifs may allow them to shift along the sequence during evolution without compromising the binding. The functional importance of these linear motifs recently prompted the development of specific alignment algorithms and dedicated benchmarks (52,53).

Cdc45 test case clearly illustrates how alignment pitfalls may hinder proper binding site prediction even with as closely related species as *Saccharomyces* ones. Our large scale analysis of Rad53 partners shows that inspection of the alignment in the vicinity of the phospho-residue may provide crucial hints to rescore the binding sites. The STRIP web server facilitates such analysis by allowing the user to analyze around each putative phospho-residue a fragment of the multiple sequence alignment with different sequence highlights (Supp Figure 4). To date, the STRIP server is dealing with *Saccharomyces* datasets and will further progress by integrating data for plants and mammals.

One major question raised by the example of the FHA1 domain is whether the most affine motif identified through peptide library screening is really useful to predict FHA1 binding motifs *in vivo*. In two well-studied interactions, Rad53 FHA1 was found to recognize its partners Rad9 and Pin4/Mdt1 through threonines within TxxV or TxxI motifs, respectively (20,23,54). Our analysis restores the reliability of the consensus motif analysis showing it could guide efficiently the predictions for the Ptc2, Cdc7 and Cdc45 examples. Presence of the pTxxD consensus motif significantly contributes to reach affinities in the range 0.5-1 μM, while the affinities of the pT motifs studied in the Rad9 and Mdt1 were an order of magnitude lower. A specificity of Rad9 and Mdt1 is to be hyperphosphorylated upon genotoxic stress by the Phosphatidyl Inositol Kinase-like Kinases (PIKKs) Tel1 and Mec1 on their SQ/TQ-rich clusters. In the case of Rad9, and probably Mdt1 also, these clusters are essential in mediating the interaction with Rad53 through its FHA domains. High phosphorylation of the SQ/TQ-rich clusters may provide multiple substrates for the FHA domains that release the stringency for a specific consensus motif with a sub-μM affinity. In case the Rad53 binding partners have more

isolated phospho-threonines, it is reasonable to think that the consensus motif rule will be more prevalent.

The molecular logic underlying phosphoproteome organizations will surely benefit from the development of STRIP-like strategies. Dissection of the intricate network of interactions between the components of cell signalling systems and/or cell machineries is all the more difficult that the redundancy of their contacts make the role of each interaction difficult to analyze (55). Systematic prediction of the contacting sites is expected to help overcoming these issues. Furthermore, competitive or synergistic interactions between interacting modules may be further predicted from the identification of the precise binding sites. A growing list of proteins involved in key signalling processes also demonstrate that alternative post-translational modifications such as acetylations or methylations may synergize with the phosphorylation of a particular site to implement complex regulatory signals (56,57). These proteins are under specific focus but such level of complexity may be more widespread and the development of predictive strategies to isolate a limited number of putative binding sites between proteins should have a major impact in the global understanding of cell components cross-talks.

References

1. Seet, B. T., Dikic, I., Zhou, M. M., and Pawson, T. (2006) Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 7, 473-483
2. Bhattacharyya, R. P., Remenyi, A., Yeh, B. J., and Lim, W. A. (2006) Domains, Motifs, and Scaffolds: The Role of Modular Interactions in the Evolution and Wiring of Cell Signaling Circuits. *Annu Rev Biochem*
3. Pawson, T., and Nash, P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science* 300, 445-452
4. Gong, W., Zhou, D., Ren, Y., Wang, Y., Zuo, Z., Shen, Y., Xiao, F., Zhu, Q., Hong, A., Zhou, X., Gao, X., and Li, T. (2008) PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. *Nucleic Acids Res* 36, D679-683
5. Ceol, A., Chatr-aryamontri, A., Santonico, E., Sacco, R., Castagnoli, L., and Cesareni, G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res* 35, D557-560
6. Diella, F., Gould, C. M., Chica, C., Via, A., and Gibson, T. J. (2008) Phospho.ELM: a database of phosphorylation sites--update 2008. *Nucleic Acids Res* 36, D240-244
7. Hofmann, K., and Bucher, P. (1995) The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem Sci* 20, 347-349
8. Liao, H., Yuan, C., Su, M. I., Yongkiettrakul, S., Qin, D., Li, H., Byeon, I. J., Pei, D., and Tsai, M. D. (2000) Structure of the FHA1 domain of yeast Rad53 and identification of binding sites for both FHA1 and its target protein Rad9. *J Mol Biol* 304, 941-951
9. Durocher, D., Taylor, I. A., Sarbassova, D., Haire, L. F., Westcott, S. L., Jackson, S. P., Smerdon, S. J., and Yaffe, M. B. (2000) The molecular basis of FHA domain:phosphopeptide binding specificity and implications for phospho-dependent signaling mechanisms. *Mol Cell* 6, 1169-1182
10. Durocher, D., Henckel, J., Fersht, A. R., and Jackson, S. P. (1999) The FHA domain is a modular phosphopeptide recognition motif. *Mol Cell* 4, 387-394
11. Li, J., Williams, B. L., Haire, L. F., Goldberg, M., Wilker, E., Durocher, D., Yaffe, M. B., Jackson, S. P., and Smerdon, S. J. (2002) Structural and functional versatility of the FHA domain in DNA-damage signaling by the tumor suppressor kinase Chk2. *Mol Cell* 9, 1045-1054
12. Huen, M. S., Grant, R., Manke, I., Minn, K., Yu, X., Yaffe, M. B., and Chen, J. (2007) RNF8 transduces the DNA-damage signal via histone ubiquitylation and checkpoint protein assembly. *Cell* 131, 901-914
13. Lee, H., Yuan, C., Hammet, A., Mahajan, A., Chen, E. S., Wu, M. R., Su, M. I., Heierhorst, J., and Tsai, M. D. (2008) Diphosphothreonine-specific interaction between an SQ/TQ cluster and an FHA domain in the Rad53-Dun1 kinase cascade. *Mol Cell* 30, 767-778
14. Koch, C. A., Agyei, R., Galicia, S., Metalnikov, P., O'Donnell, P., Starostine, A., Weinfeld, M., and Durocher, D. (2004) Xrcc4 physically links DNA end processing by polynucleotide kinase to DNA ligation by DNA ligase IV. *Embo J* 23, 3874-3885
15. Byeon, I. J., Li, H., Song, H., Gronenborn, A. M., and Tsai, M. D. (2005) Sequential phosphorylation and multisite interactions characterize specific target recognition by the FHA domain of Ki67. *Nat Struct Mol Biol* 12, 987-993
16. Yuan, C., Yongkiettrakul, S., Byeon, I. J., Zhou, S., and Tsai, M. D. (2001) Solution structures of two FHA1-phosphothreonine peptide complexes provide insight into the structural basis of the ligand specificity of FHA1 from yeast Rad53. *J Mol Biol* 314, 563-575
17. Tam, A. T., Pike, B. L., and Heierhorst, J. (2008) Location-specific functions of the two forkhead-associated domains in Rad53 checkpoint kinase signaling. *Biochemistry* 47, 3912-3916
18. Smolka, M. B., Chen, S. H., Maddox, P. S., Enserink, J. M., Albuquerque, C. P., Wei, X. X., Desai, A., Kolodner, R. D., and Zhou, H. (2006) An FHA domain-mediated protein interaction network of Rad53 reveals its role in polarized cell growth. *J Cell Biol* 175, 743-753
19. Schwartz, M. F., Lee, S. J., Duong, J. K., Eminaga, S., and Stern, D. F. (2003) FHA domain-mediated DNA checkpoint regulation of Rad53. *Cell Cycle* 2, 384-396
20. Pike, B. L., Yongkiettrakul, S., Tsai, M. D., and Heierhorst, J. (2004) Mdt1, a novel Rad53 FHA1 domain-interacting protein, modulates DNA damage tolerance and G(2)/M cell cycle progression in Saccharomyces cerevisiae. *Mol Cell Biol* 24, 2779-2788
21. Pike, B. L., Yongkiettrakul, S., Tsai, M. D., and Heierhorst, J. (2003) Diverse but overlapping functions of the two forkhead-associated (FHA) domains in Rad53 checkpoint kinase activation. *J Biol Chem* 278, 30421-30424

22. Pike, B. L., Hammet, A., and Heierhorst, J. (2001) Role of the N-terminal forkhead-associated domain in the cell cycle checkpoint function of the Rad53 kinase. *J Biol Chem* 276, 14019-14026
23. Mahajan, A., Yuan, C., Pike, B. L., Heierhorst, J., Chang, C. F., and Tsai, M. D. (2005) FHA domain-ligand interactions: importance of integrating chemical and biological approaches. *J Am Chem Soc* 127, 14572-14573
24. Duncker, B. P., Shimada, K., Tsai-Pflugfelder, M., Pasero, P., and Gasser, S. M. (2002) An N-terminal domain of Dbf4p mediates interaction with both origin recognition complex (ORC) and Rad53p and can deregulate late origin firing. *Proc Natl Acad Sci U S A* 99, 16087-16092
25. Bjergbaek, L., Cobb, J. A., Tsai-Pflugfelder, M., and Gasser, S. M. (2005) Mechanistically distinct roles for Sgs1p in checkpoint activation and replication fork maintenance. *Embo J* 24, 405-417
26. Fromont-Racine, M., Rain, J. C., and Legrain, P. (2002) Building protein-protein networks by two-hybrid mating strategy. *Methods Enzymol* 350, 513-524
27. Guillemain, G., Ma, E., Mauger, S., Miron, S., Thai, R., Guerois, R., Ochsenbein, F., and Marsolier-Kergoat, M. C. (2007) Mechanisms of checkpoint kinase Rad53 inactivation after a double-strand break in Saccharomyces cerevisiae. *Mol Cell Biol* 27, 3378-3389
28. Blom, N., Gammeltoft, S., and Brunak, S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol* 294, 1351-1362
29. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., and Dunker, A. K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res* 32, 1037-1049
30. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680
31. Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* 301, 71-76
32. Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423, 241-254
33. Leroy, C., Lee, S. E., Vaze, M. B., Ochsenbien, F., Guerois, R., Haber, J. E., and Marsolier-Kergoat, M. C. (2003) PP2C phosphatases Ptc2 and Ptc3 are required for DNA checkpoint inactivation after a double-strand break. *Mol Cell* 11, 827-835
34. Kihara, M., Nakai, W., Asano, S., Suzuki, A., Kitada, K., Kawasaki, Y., Johnston, L. H., and Sugino, A. (2000) Characterization of the yeast Cdc7p/Dbf4p complex purified from insect cells. Its protein kinase activity is regulated by Rad53p. *J Biol Chem* 275, 35051-35062
35. Xue, Y., Zhou, F., Zhu, M., Ahmed, K., Chen, G., and Yao, X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res* 33, W184-187
36. Huang, H. D., Lee, T. Y., Tzeng, S. W., and Horng, J. T. (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res* 33, W226-229
37. Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics* 4, 1633-1649
38. Xue, Y., Li, A., Wang, L., Feng, H., and Yao, X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics* 7, 163
39. Obenauer, J. C., Cantley, L. C., and Yaffe, M. B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* 31, 3635-3641
40. Beltrao, P., and Serrano, L. (2007) Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol* 3, e25
41. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636
42. Herzberg, K., Bashkirov, V. I., Rolfsmeier, M., Haghnazari, E., McDonald, W. H., Anderson, S., Bashkirova, E. V., Yates, J. R., 3rd, and Heyer, W. D. (2006) Phosphorylation of Rad55 on serines 2, 8, and 14 is required for efficient homologous recombination in the recovery of stalled replication forks. *Mol Cell Biol* 26, 8396-8409
43. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* 415, 180-183
44. Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Hofert, C., et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147

45. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* 440, 637-643
46. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34, D535-539
47. Breitkreutz, B. J., Stark, C., and Tyers, M. (2003) Osprey: a network visualization system. *Genome Biol* 4, R22
48. Smolka, M. B., Albuquerque, C. P., Chen, S. H., and Zhou, H. (2007) Proteome-wide identification of in vivo targets of DNA damage checkpoint kinases. *Proc Natl Acad Sci U S A* 104, 10364-10369
49. Sidorova, J. M., and Breeden, L. L. (2003) Rad53 checkpoint kinase phosphorylation site preference identified in the Swi6 protein of Saccharomyces cerevisiae. *Mol Cell Biol* 23, 3405-3416
50. Rudra, D., Mallick, J., Zhao, Y., and Warner, J. R. (2007) Potential interface between ribosomal protein production and pre-rRNA processing. *Mol Cell Biol* 27, 4815-4824
51. Betel, D., Breitkreuz, K. E., Isserlin, R., Dewar-Darch, D., Tyers, M., and Hogue, C. W. (2007) Structure-templated predictions of novel protein interactions from sequence information. *PLoS Comput Biol* 3, 1783-1789
52. Chica, C., Labarga, A., Gould, C. M., Lopez, R., and Gibson, T. J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics* 9, 229
53. Perrodou, E., Chica, C., Poch, O., Gibson, T. J., and Thompson, J. D. (2008) A new protein linear motif benchmark for multiple sequence alignment software. *BMC Bioinformatics* 9, 213
54. Schwartz, M. F., Duong, J. K., Sun, Z., Morrow, J. S., Pradhan, D., and Stern, D. F. (2002) Rad9 phosphorylation sites couple Rad53 to the Saccharomyces cerevisiae DNA damage checkpoint. *Mol Cell* 9, 1055-1065
55. Palmbos, P. L., Daley, J. M., and Wilson, T. E. (2005) Mutations of the Yku80 C terminus and Xrs2 FHA domain specifically block yeast nonhomologous end joining. *Mol Cell Biol* 25, 10782-10790
56. Yang, X. J. (2005) Multisite protein modification and intramolecular signaling. *Oncogene* 24, 1653-1662
57. Calnan, D. R., and Brunet, A. (2008) The FoxO code. *Oncogene* 27, 2276-2288

Aknowledgements

W. A. was financed by an ACI IMPBIO grant. This work was financed in part by the Association pour la Recherche sur le Cancer and by an ANR grant.

Figure legends

Figure 1. (A) Cdc7 threonine STRIP analysis. (B) Cdc45 threonine STRIP analysis. Threonines in bold and italics correspond to those present inside and outside the interacting fragments identified by the two-hybrid screens, respectively.

Figure 2. Two-hybrid assay monitoring the interactions between FHA1 [Rad53(1-164)] and either Cdc7(294-493) (A) or Cdc45(154-270) (B). pGBT9/FHA1 or pGBT9/FHA1R70A expressing the Gal4BD-Rad53(1-164) (wt) and Gal4BD-Rad53(1-164)R70A (m) fusion proteins, respectively, were introduced into the tester strain along with wild-type or mutated pACTII/Cdc7(294-493) or pACTII/Cdc45(154-270) vectors harboring the sequence encoding Gal4AD fused to the sequences encoding wild-type (WT) Cdc7(294-493) or Cdc45(154-270), or their mutated derivatives as indicated. - indicates that empty vectors (either pGBT9 or pACTII) were used as controls. The two-hybrid interactions were revealed by growth on plates lacking histidine (-His) complemented with 100 mM 3-amino-triazole (3AT) and by X-gal staining. Control plates contained 3AT but were complemented with histidine (+His). It has to be noted that the Gal4BD-Rad53(1-164) fusion protein can activate by itself the transcription of the reporter genes at a low level, hence the residual growth and the slight blue coloration visible on the first spot of the plates (for transformants containing pGBT9/FHA1 and the empty vector pACTII).

Figure 3.

Affinity between the FHA1 domain of Rad53 and phosphopetides from (A) Ptc2, (B) Cdc7 and (C) Cdc45 measured by isothermal titration calorimetry. Fitted Kd values are presented in Table 2.

Figure 4. Graph summarizing the proteins found to physically interact with Rad53 in affinity-based (blue links) and two-hybrid (red links) experiments as collected in the Biogrid database. Color codes refer to the Gene Ontology definitions used in the Osprey visualization program. Cdc7 and Cdc45 gene names are colored red and the other gene names which were studied in this work are labelled with an obelisk (‡). Proteins found to specifically interact with the isolated FHA1 domain of Rad53 through affinity-based experiments are labelled with an asterisk (*) (18). When possible, gene names were clustered with respect to their connectivity and to their biological function in black boxes. Plain or dashed underlines indicate proteins bearing a putative FHA1 binding phosphothreonine as detected using the stringent or the more permissive predictive criteria, respectively.

Supp Figure 1.

Cdc45 multiple sequence alignment in the vicinity of threonines T147 and T189. The region spanning T189 corresponds to an acidic stretch, structurally disordered, containing gaps and repetitive sequences emphasizing the difficulties of obtaining a correct alignment.

Supp Figure 2.

Bud4, Shs1 and Cdc11 multiple sequence alignments in the vicinity of the candidate binding threonines. The three proteins are the only members of the septin complex to bear threonines likely to be bound by the FHA1 domain of Rad53.

Supp Figure 3

Swi4 and Swi6 multiple sequence alignments in the vicinity of the candidate binding threonines. Both proteins are the only members of the SWI complex to bear threonines likely to be bound by the FHA1 domain of Rad53.

Supp Figure 4

Screen capture of the STRIP web server with the query page on the left and the result page on the right. A dark blue color indicates that the corresponding phosphoresidue satisfies either the phosphorylability, the consensus motif rule or the strict conservation condition. Light blue cells indicate that the conservation condition is only respected with the less stringent condition and in that case, further manual analysis may be performed easily. By clicking in the table cell of any phosphoresidue, a pop-up window allows for rapidly checking the conservation features between *S. cerevisiae* close homologs, detecting possible sequence truncations or identifying alignment flaws.

Table 1.
FHA1 interactants identified throught the two-hybrid screenings. 'Protein fragment' corresponds to the shortest protein fragment that was found interacting with FHA1. 'Clones' indicates the number of independent clones that were recovered among the hits. +/- CPT refers to the screenings that were performed either in the absence of genotoxic stress or in the presence of 5 μg/ml camptothecin (CPT).

ORF name	Protein fragment	Hits - CPT	Clones - CPT	Hits + CPT	Clones + CPT	Description	Reference
YNL084C	End3(262-349)	16	7	4	4	Protein involved in endocytosis, actin cytoskeletal organization and cell wall morphogenesis	
YLR103C	Cdc45(154-270)	5	3	11	4	DNA replication initiation factor	
YDL017W	Cdc7(294-493)	0	0	4	4	Dbf4-dependent kinase catalytic subunit required for firing origins and replication fork progression	(34)
YFR022W	Rog3(383-604)	4	2	5	2	Protein that binds to Rsp5, a ubiquitin ligase; overexpression of RSP5 rescues the end3D mutant	
YGR013W	Snu71(173-359)	2	2	1	1	Component of U1 snRNP required for mRNA splicing	
YJR031C	Gea1(1-292)	1	1	2	1	Guanine nucleotide exchange factor for ADP ribosylation factors (ARFs), involved in vesicular transport between the Golgi and ER, Golgi organization, and actin cytoskeleton organization	
YPL120W	Vps30(104-190)	4	2	0	0	Protein that forms a membrane-associated complex essential for autophagy; involved in vacuolar protein sorting	
YOL028C	Yap7(44-245)	1	1	1	1	bZIP transcription factor	(18)
YML016C	Ppz1(158-443)	1	1	1	1	S/T phosphatase	
YHR202W	Yhr202w(498-602)	1	1	1	1	Putative protein of unknown function	
YER089C	Ptc2(61-456)	1	1	1	1	PP2C phosphatase	(33)

Table 2
Thermodynamic parameters for FHA1 binding to Ptc2 (DDIpTDADTDAE), Cdc7 (DGESpTDEDDVVS) and Cdc45 (DDEApTDADEVTD) phosphopeptides measured by isothermal titration calorimetry (ITC)

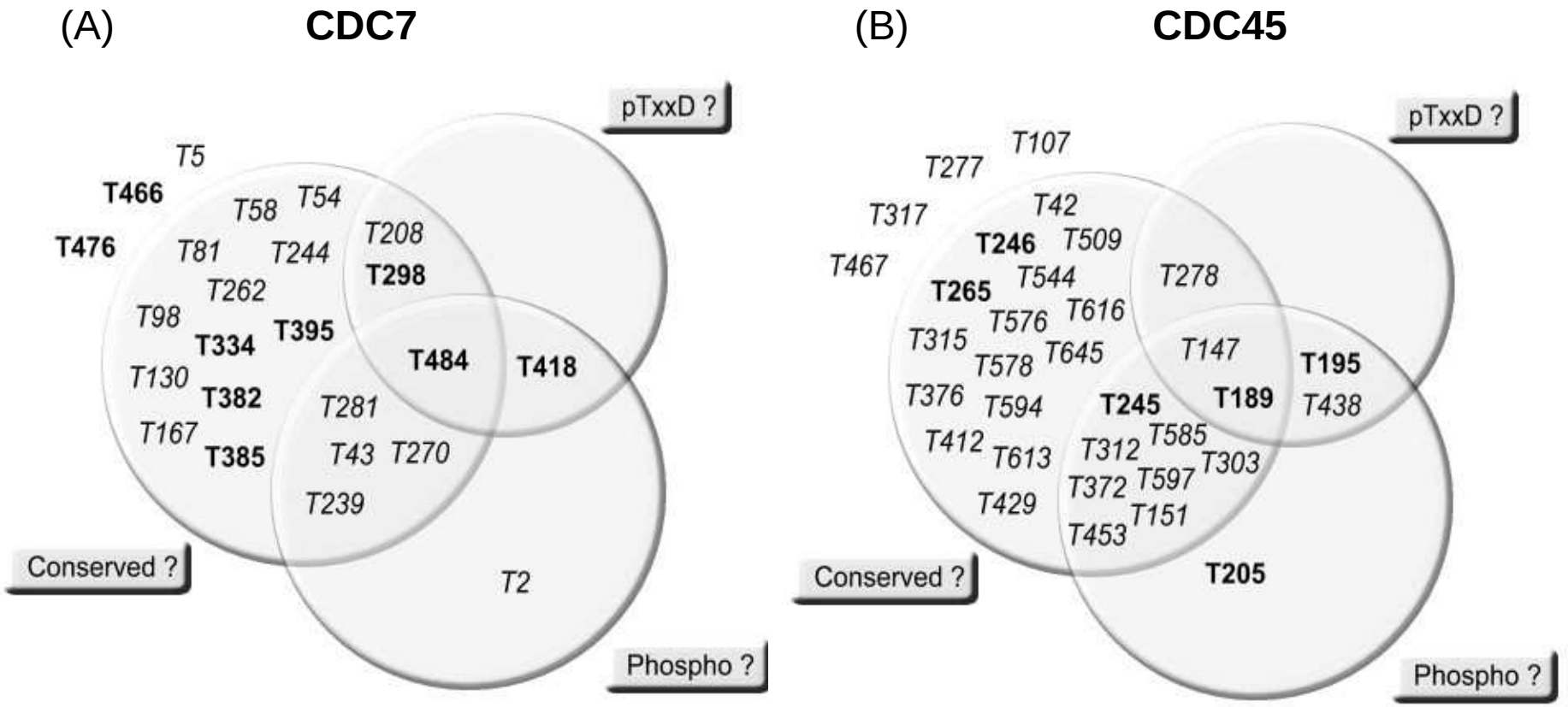
peptide	stoichiometry	K_d (μM)	ΔH (kcal/mol)	ΔG (kcal/mol)	$-T\Delta S$ (kcal/mol)
pT-Ptc2	1	0.7±0.01	-21.5±0.05	-8.5	-13.0
pT-Cdc7	1	1.69±0.04	-20.6±0.10	-8.0	-12.6
pT-Cdc45	0.97	0.42±0.02	-20.7±0.10	-8.8	-11.9

Table 3 : STRIP-based prediction of the threonines possibly involved in the interaction with the FHA1 domain of Rad53. (‡) gene names of the proteins identified in the FHA1 two-hybrid screen performed in this work. Predictions were carried out using the full-length proteins. (#) label the predicted threonines that are comprised within the fragment identified in the two-hybrid screen performed in this work. (*) proteins that were identified in the affinity-based screen by Smolka and co-workers using the isolated FHA1 as a bait (18).

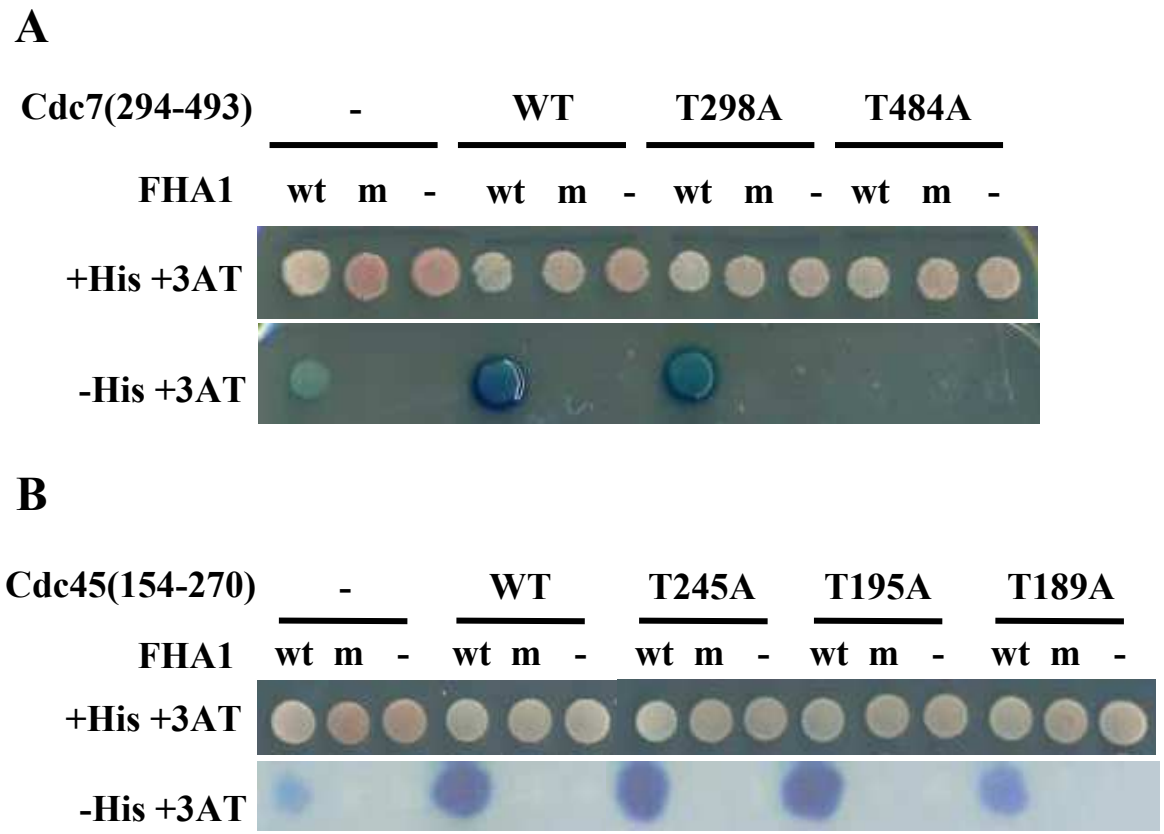
	Number of Threonines	Number of TxxD motifs	Threonines strict conditions	Threonines permissive conditions
End3‡	11	0		
Rog3‡	33	2		T453#
Snu71‡	29	3	T269#	
Gea1‡	76	4		T272#
Yap7‡	10	0		
Vps30‡	36	6	T434	T130#; T257
Ppz1‡	38	2	T171#	
Yhr202w‡	38	1	T504#	
Cdc7‡	24	5	T484	
Dbf4	68	5		T247 ; T253
Mrc1*	69	7		T242 ; T272 ; T977
Cdc45‡	33	5		T147 ; T189
Cdc46	49	1		
Rad9*	89	5		
Ptc2*‡	32	4	T376#	
Dun1	26	0		
Mec1	131	10		
Hhf1	6	1		
Hht1	9	0		
Hta2	5	0		
Hmo1	13	0		
Asf1*	9	1	T270	
Swi6*	41	2		T111
Mbp1*	63	4		
Swi4*	70	3		T64
Whi5*	39	0		
Kap95	44	1		
Srp1/Kap60	32	2		T273
Gln3*	49	2		
Ifh1*	55	7	T346	

Tbf1	42	1		T522
Cdc13	54	1		
Cst6*	48	2	T534	
Esc1*	98	2		
Sgs1	103	3	T423	
Mus81	81	2		
Rad55	22	1		
Crp1*	38	2		T170
Src1*	36	3		T399
Ecm16*	73	5	T869 ; T1169	
Net1*	82	2		
Yta7*	70	7	T247; T946 ; T1077	
Psy2	59	7		T12 ; T178 ; T524
Smc3	66	5	T18	
Cdc3*	18	1		
Cdc12*	27	0		
Shs1*	36	2		T539
Cdc11*	30	1		T62
Cdc10*	24	1		
Bud4*	88	8		T178 ; T237 ; T612
Bud3*	106	7	T255	
Bud14*	36	2		
Bmh2	15	1		T4
Sec2*	52	5		T423
Ubp1*	37	4		T592
Rck2*	39	4		T342 ; T539
Fyv8*	50	7	T602; T608	
Mnr2*	52	5		
Ipp1	19	2	T100 ; T247	
Ckb2	10	0		
Ede1	118	3		T948
Pin4	39	0		
Gid8	27	0		

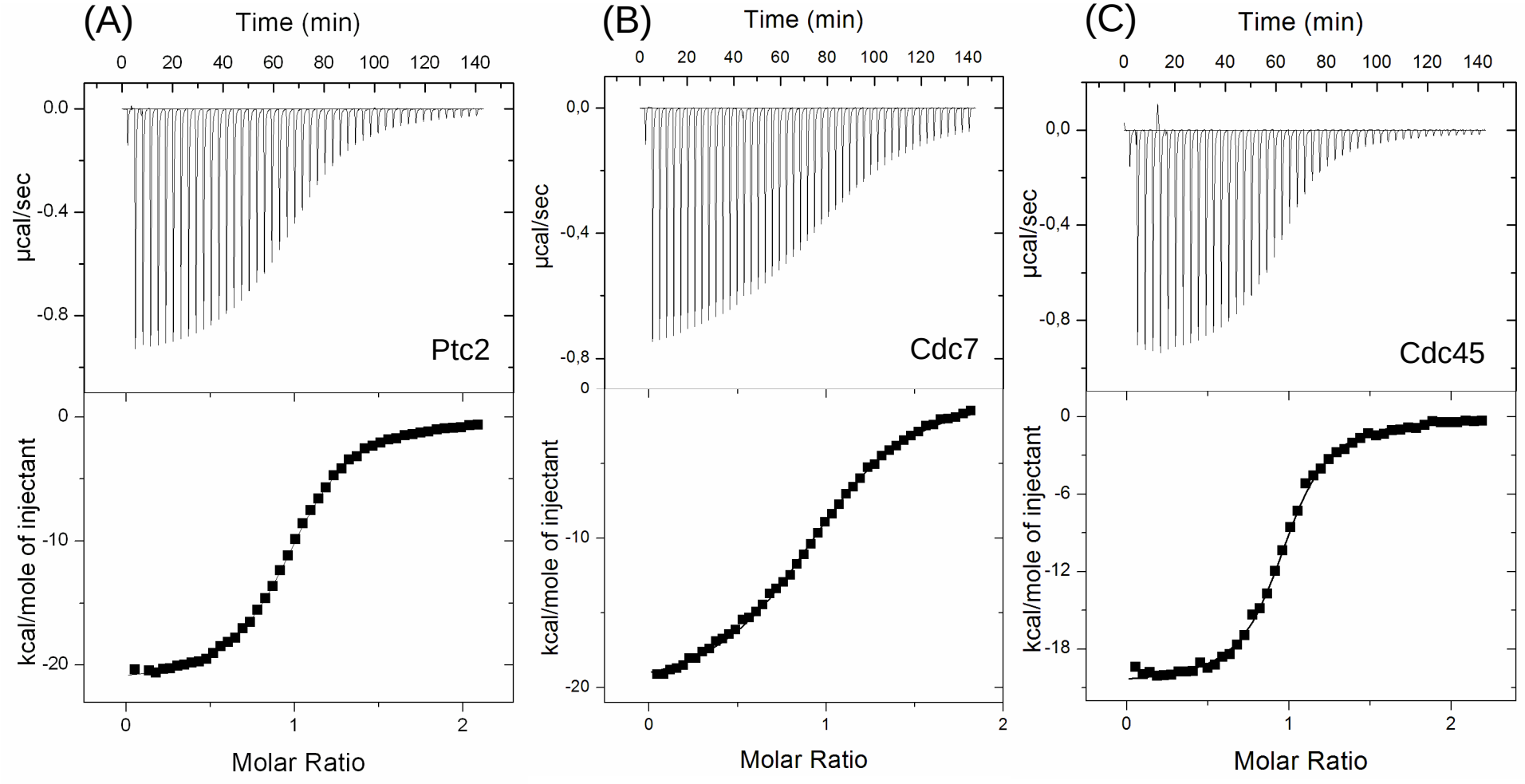
Aucher_etal_Figure 1



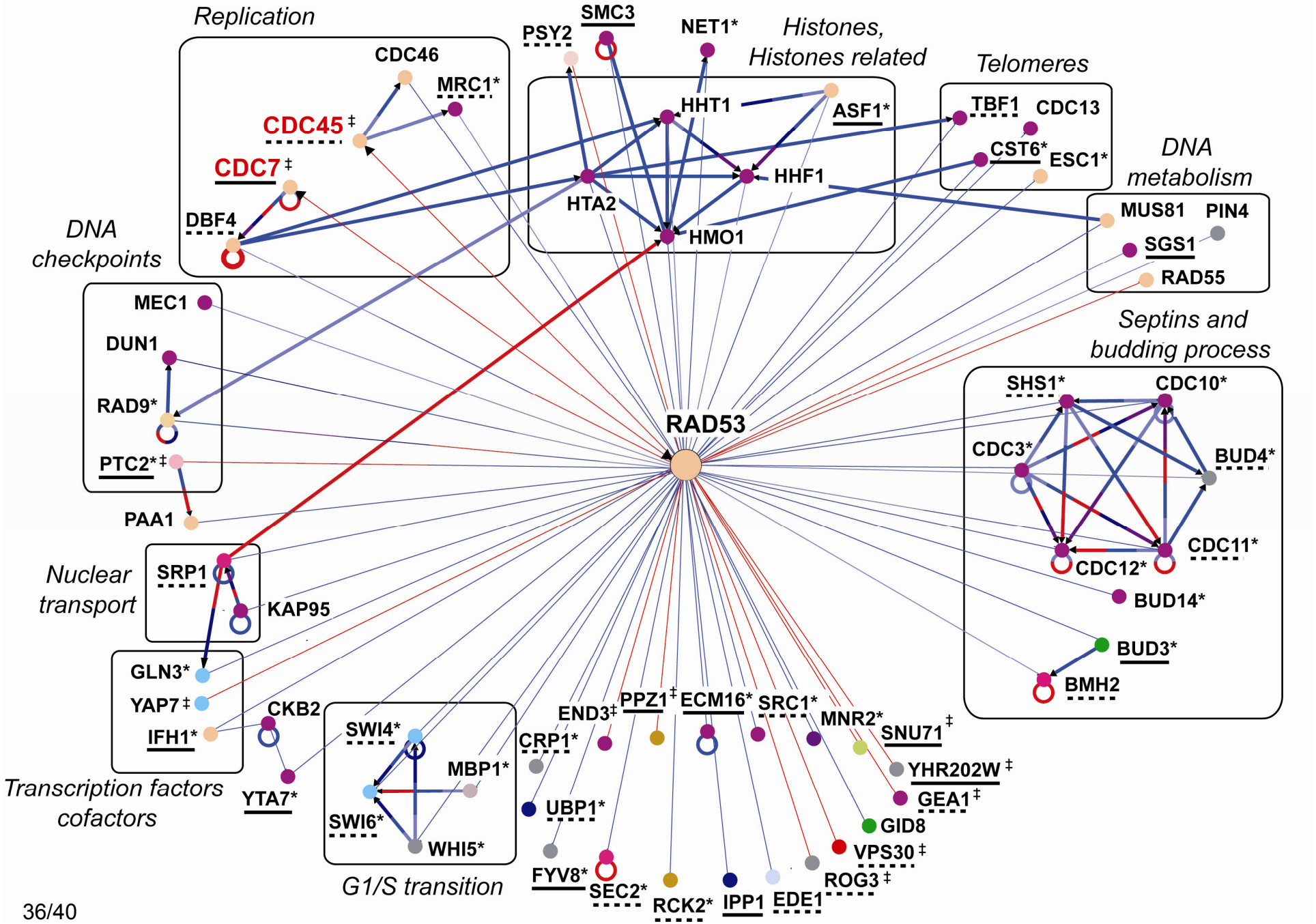
Aucher_etal_Figure 2



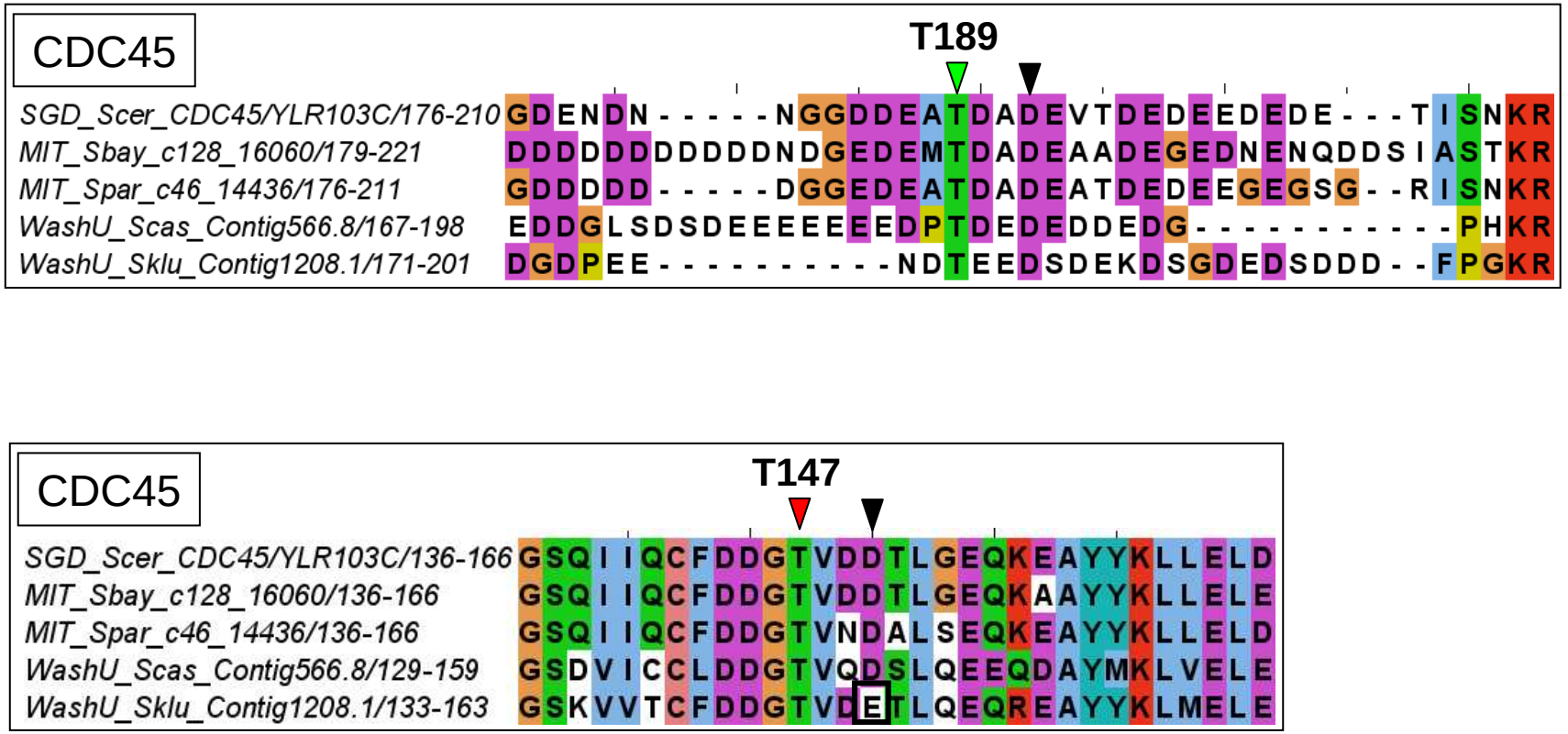
Aucher_etal_Figure 3



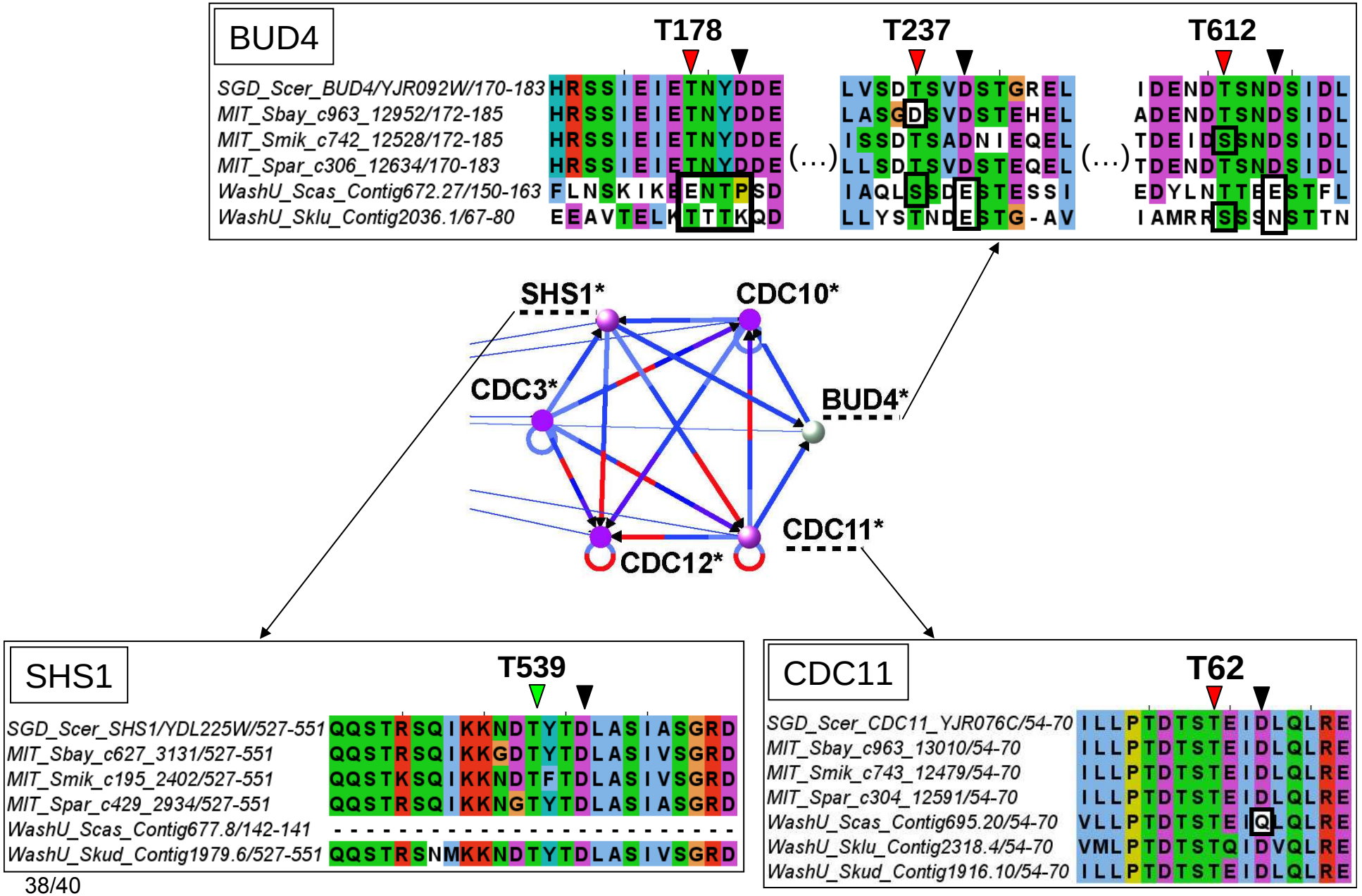
Aucher_etal_Figure 4



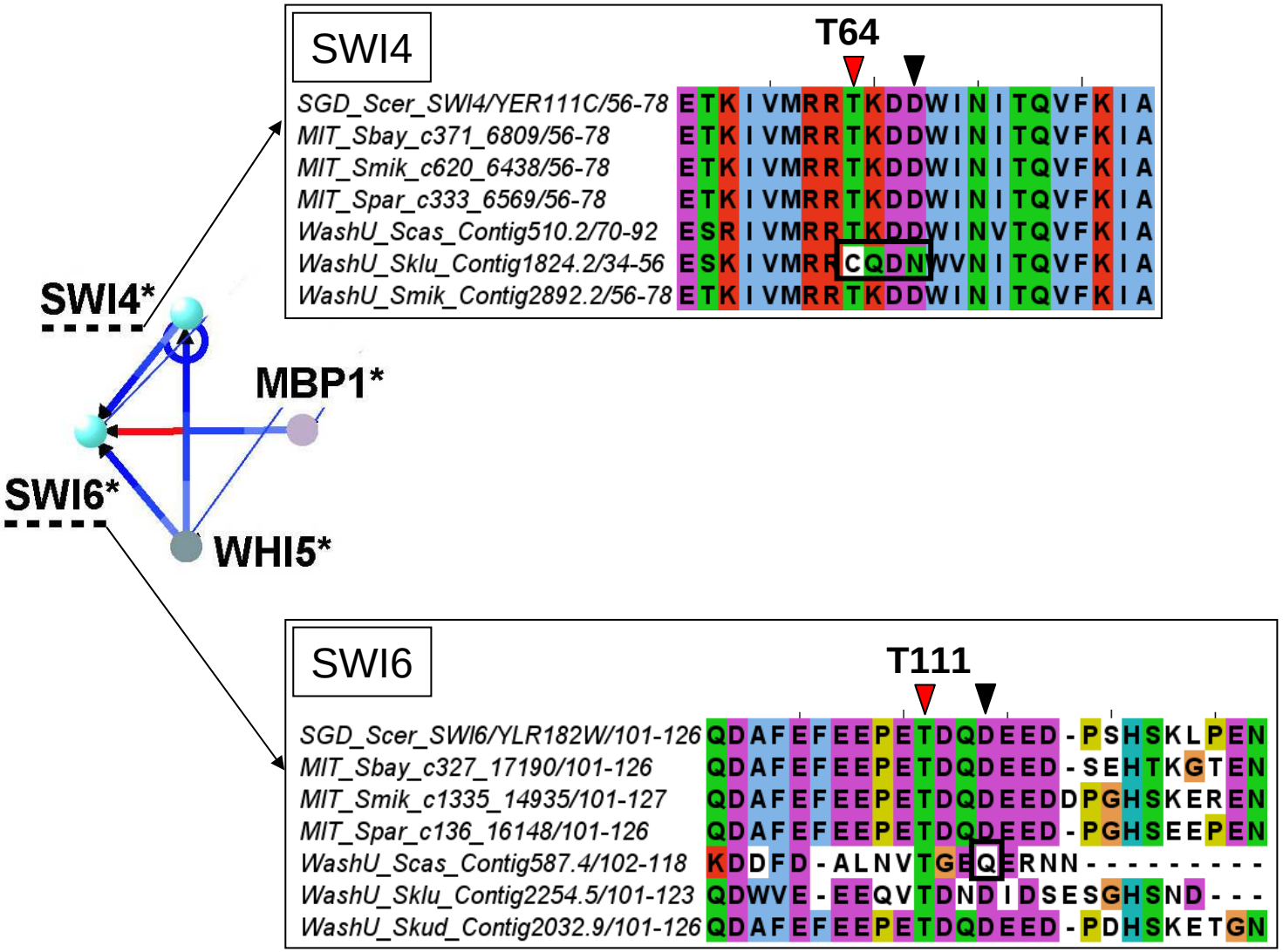
Aucher_etal_Supp Figure 1



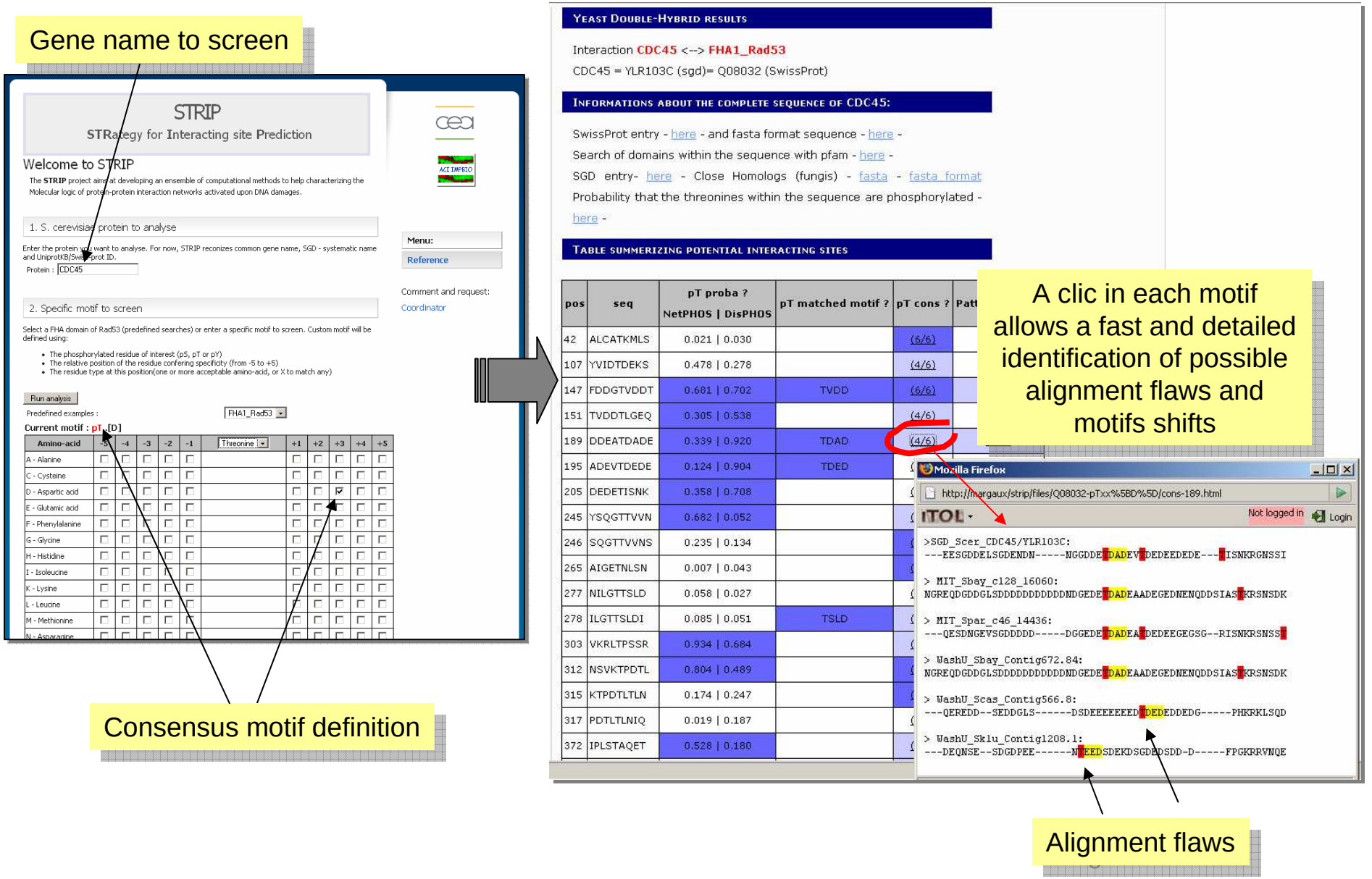
Aucher_etal_Supp Figure 2



Aucher_etal_Supp Figure 3



Aucher_etal_Supp Figure 4



Les protéines impliquées dans les voies de signalisation sont souvent activées et inactivées par des interactions de faible affinité. En particulier, les domaines protéiques liant spécifiquement de courts fragments protéiques permettent une régulation intra- et inter-moléculaire efficace des domaines catalytiques auxquels ils sont associés. Citons par exemple les domaines FHA ou des tandems BRCT fréquemment impliqués dans les réponses aux dommages de l'ADN. Etant donnée leur importance dans les réseaux d'interactions et dans la signalisation cellulaire, la prédiction par bioinformatique des propriétés de liaison de ces petits domaines constitue un enjeu majeur. Toutefois, les stratégies bioinformatiques sont jusqu'à présent limitées par des difficultés méthodologiques associées aux caractéristiques intrinsèques de ces domaines. Leurs séquences sont souvent très divergentes et les affinités pour leurs cibles physiologiques sont généralement faibles malgré une excellente spécificité. Le travail présenté dans cette thèse a donc pour objectif de dépasser les limites actuelles des outils de prédictions pour développer de nouvelles méthodologies bioinformatiques performantes. Trois points ont été plus particulièrement abordés : (i) la prédiction de la structure tridimensionnelle de ces domaines ; (ii) la prédiction des sites reconnus par ces domaines lorsque les partenaires sont connus ; (iii) la prédiction des motifs spécifiquement reconnus par ces domaines sur la base de leur structure tridimensionnelle.

Proteins involved in signalling pathways are frequently activated and inactivated by weak affinity interactions. In particular, domains that bind specifically short protein fragments, often called Peptide Recognition Modules (PRMs), allow an efficient intra- and inter-molecular regulation of the catalytic domains to whom they are associated. This work focuses on two domain families, the FHA and the tandem BRCT, often involved in the cell responses to DNA damage. Given their major role in the signaling networks, the prediction of their binding properties by bioinformatics is of crucial interest. However, bioinformatic strategies are still limited by methodological problems associated with the intrinsic characteristics of PRMs. They typically harbour very divergent sequences and their affinity for their physiological target is relatively weak despite their high specificity. This work aims at developing predictive approaches for the study of PRMs. Three points have been considered successively : (i) the prediction of the three-dimensional structure of PRMs, (ii) the prediction of their binding sites when the partner is known, (iii) the prediction of the sequence motifs each PRM specifically recognizes based on its three-dimensional structure.